University of California
Santa Barbara

Peripheral Representations: from Perception to Visual Search

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Dynamical Neuroscience

by

Manuel Arturo Deza Figueroa

Committee in charge:

Professor Miguel P. Eckstein, Chair
Professor Barry Giesbrecht
Professor B.S. Manjunath

March 2019

The Dissertation of Manuel Arturo Deza Figueroa is approved.

Professor Barry Giesbrecht

Professor B.S. Manjunath

Professor Miguel P. Eckstein, Committee Chair

December 2018

Peripheral Representations: from Perception to Visual Search

by

Manuel Arturo Deza Figueroa

A mi patria,

Acknowledgements

This PhD thesis could not have been finished if it had not been due to all the supportive people I have met in my life. Including all of them might as well result in another PhD Thesis. In particular I would like to thank and dedicate this piece of work to the following people mentioned in this section.

To my advisor and friend Miguel, who took a great risk in taking me as a student. I had joined the lab with a background in Robotics and had no idea of what the visual cortex was (I still feel like I don't), and Miguel slowly enriched my hacker persona into a scientist. I am indebted to him for seeing my drive and motivation when I was still an undergraduate applicant from Peru who happened to sit in his vision science class out of mere curiosity. I am grateful that in these five years he has helped me become more of a vision scientist than I could have ever imagined. I will always remember his playful signature chant that he delivered with a big smile on his face during lunch: *"Che, yo quiero Deep Understanding!"*. Indeed, I do not think I would have embraced the freedom of work, rigorous evaluation, and wide variety of projects had I been in another lab anywhere else in the world.

To Manjunath who gave me the first opportunity to prove myself as an intern at UCSB back in 2012. It was at his lab, the Center for BioImage Informatics, where I started my first steps in image processing, under the close guidance of Aruna (his graduate student, my mentor and friend!). It was at this time that I met Devi Parikh who later accepted me as a Computer Vision intern at Virginia Tech to pursue a project on image virality. I owe her teaching me the critical elements on how to get papers written and published in computer vision venues. Devi and Dhruv were starting an amazing Computer Vision and Machine Learning group at Virginia Tech, and I felt privileged to witness how that lab started from its beginnings to where they are now. It was also through working with her

during that short stay, that I met many friends that I still keep in touch with today such as Senthil and Abhimanyu, which I've continued to see at computer vision and machine learning conferences.

I'd also like to thank Barry who always seemed to find specific question in my qualifying exam that would make me re-evaluate how little I knew about things I thought I had under my belt, and for taking the time to read my thesis and point out critical aspects of re-organization and flow that were missing in preliminary drafts. I felt that Barry was an amazing shadow advisor who always seemed to know when I was going in the right direction, or when I was taking a wrong turn, and I greatly appreciate his guidance. Craig was always open to talk about math when I needed some orientation on a problem I was working on, and gave me the most critical feedback for the metamer project – that in hindsight would have not succeeded had it not been for his thorough evaluation.

I'd also like to give a special thank you to Amit, who played a pivotal role as a mentor and friend over a long 4 year collaboration in much of the work performed in this thesis. He was almost like a secondary advisor, and I owe him the opportunity to have joined United Techologies Resarch Center (UTRC) as a computer vision and deep learning intern to experience the rapid life of industry with Jeff, Kishore and Edgar.

My high school friends: there are too many stories to write and share, I can't thank you all enough to have shaped me as an outgoing Peruvian, compared to the relative quite and shy person I was when I just joined middle school moving back from America. You have all taught me that there is more in life than just work: Gonzalo, Manuel, Giancarlo, Francisco, Alejandro, Percy, Rodrigo, Andre. The brotherhood that we developed over my time with you all in Peru will never go away, as it will always be the place that I call home.

I will also never forget the early days in college where I would fantasize about leaving

the country and finding a better life back in America. Luis Alonso, Fernando, Alex, Juan Manuel, Luis Carlos, Juan Elmer, Celso the conversations we had when would sit in the table at the library (BiguFIM), or over lunch, or on the bus, and would talk about politics, dreams, homework and the many frustrations that we had to go through every day where life seemed to be moving nowhere. All my friends from GISCIA who allowed me to explore research in the most creative and unrestricted way: Renato, Jorge, Oscar, David, Kevin, Jhair, Alexander and Alejandro. We were a group of undergraduates who would get together and work on projects we found interesting – and we did it just for fun, without knowing how it would shape our future. The amazing teachers I had from my college, there are too many of you to mention, but you all contributed in great ways by making college classes seem impossible. In retrospect, my formative years at Universidad Nacional de Ingenieria (*a.k.a* The Prison of Blue Walls) trained me with the stamina and mental endurance, that were necessary to prepare me in my path to becoming a scientist later in life at the University of California, Santa Barbara (The Paradise of Blue Walls).

My roommates Christian, Garo and Mihir who made my life feel like I was in a permanent sit-com. The great conversations, the parties, the anecdotes, the fights, and the brotherhood, we had was unmatchable for those 4 years that we shared an apartment together. I'm thankful that Christian was happy to teach me about differential geometry and topology in his free time, Garo seemed to have looked up to me and implicitly pushed me to be a better version of myself in everything I did, and Mihir made me feel like I was never working hard enough, as he would always walk in an hour later than I did after a long day at the lab. I don't think I would have achieved this milestone had it not been for the incredible support we all gave each other in those years. Ah yes, those were the good old-days at #1229.

My buddy Henrique for the many surf sessions in the afternoon at Sands, that naturally turned into competitions of who could get the biggest wave (and who would ride

the shortest board). Nikita, Clint and Billy who were always there for a beer, lunch, hanging out, and epic trips to Peru. Ekta for always volunteering as Reviewer #2 in the papers I worked on. Zeynep and Jorge for offering me their company, and a place to crash whenever I was in LA and had to travel.

Everyone at the VIU-Lab: Matt, Steve, Katie, Eamon, Emre, Mordechai (my buddy and financial advisor) Charles, Amir, Lauren, Miguel Lago, Aditya, Luke, Puneeth, Yuliy, Nicole, Devi and Sudhanshu. Including the amazing stories, anecdotes, trips, voyages, reunions and discussions in this section would result in a chapter by itself, how lucky am I to have spent time with you all and learned from you in ways you have not imagined. You all made being in the lab feel like we were a band of brothers, and it set the tone of the type of work environment I wanted to create and be in the future. I'd also like to thank the amazing army of Research Assistants I worked with: Alex, Jamie, Adam, Michael, Jasmine, Corey, Lenet, Juan, Kelsey, Neal, Zack, Ashi, Alexis, Zhaoqi; had it not been for you all, none of this thesis would have existed as you helped me collect the data through countless hours of the day and night.

My two college mentors Alberto and Elizabeth, who were amazing scientists, engineers, and individuals. They were very supportive in my early years as an undergraduate in Lima, Peru back when I would go visit them for weekly chats at their lab secretly called *'El Bunker'*. They were the first professors I had who encouraged me to pursue a PhD, informed me that there was a thing called graduate school, and somehow thought that I was the man for the job. Thank you for pushing me to pursue my dreams!

This thesis is finally dedicated to my family, specially my parents for always offering their unconditional love and support, and to my brother for continually believing in me – even more than I ever will.

"Waves are not measured in feet and inches, they are measured in increments of fear"

– Buzzy Trent, big wave surfer

Curriculum Vitæ

Manuel Arturo Deza Figueroa

Education

2018	Ph.D. in Dynamical Neuroscience, University of California, Santa Barbara.
2012	B.S. in Mechatronics Engineering, Universidad Nacional de Ingenieria.

Publications

Deza, A., Jonnalagadda, A., Eckstein, M.P. "Towards Metamerism via Foveated Style Transfer", *International Conference on Learning Representations (ICLR)*. 2019.

Deza, A., Surana, A., Eckstein, M.P. "Assessment of Faster R-CNN for Man-Machine collaborative search", *Submitted to IEEE Computer Vision and Pattern Recognition (CVPR) conference*. 2019.

Deza, A., Peters, J., Taylor, G.S., Surana, A., Eckstein, M.P. "Attention Allocation Aid for Visual Search", *ACM Conference on Human Factors in Computing Systems (CHI)*, 2017.

Deza, A., Eckstein, M.P. "Can Peripheral Representations Improve Clutter Metrics on Complex Scenes?", *Neural Information Processing Systems (NIPS)*, Barcelona, Spain, December, 2016.

Deza, A., Parikh, D. "Understanding Image Virality", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA. June, 2015.

Abstract

Peripheral Representations: from Perception to Visual Search

by

Manuel Arturo Deza Figueroa

The human visual field is composed of a high acuity region at the center of gaze called the fovea, and its complement, the visual periphery. Not much is known about the computations and representations of the visual periphery, as most of the focus in the field of human (and machine) vision is geared towards foveal vision. Thus, the focus of this thesis will be on understanding the computations performed by the human visual system in the visual periphery. In doing so, I will begin by modelling the perception of clutter and how it changes as a function of the behavioural task and point of fixation, developing a collection of foveated clutter models that enhance non-foveated models. I will then propose a new metamer model that renders how the information is distorted in the visual field, and what this tells us about the computations done in the visual periphery. Finally, I will conclude with the design of two hybrid man-machine collaborative visual search systems that try to overcome the limitations in human visual search imposed by the visual periphery and observer inefficiencies in terminating exploration.

Contents

Introduction

The thesis relates to visual processing in the visual periphery. In particular the goals are to model how the visual periphery is represented in the human visual system, understand its impact on the influences of clutter on visual search, and evaluate methods to compensate for its impact and improve visual search performance/efficiency.

We begin this thesis with the study of image complexity in the context of visual clutter in the first chapter. As the visual field is spatially variant contingent on a point of fixation, features pool or 'crowd' stronger as a function of retinal eccentricity. We thus tested a collection of standard (non-foveated) clutter models and developed foveated version for each one of these models, to enrich their representations as a function of point of fixation. These foveated models are created by computing dense representations of each clutter model and stacking a peripheral architecture that simulates the human visual field and the early effects of pooling and retinal ganglion cells convergence from LGN to V1. We show that foveated models better predict human search for targets in scenes and clutter judgments than non-foveated models.

In the second chapter we develop a generative model that results in images that match the loss of information via distortions generated in the visual periphery. Generative models of foveated visual perception have been dubbed as visual metamers: two perceptually indistinguishable images that are physically different contingent on a point of fixation [1]. An example of such metamers can be seen in Figure 0.1. Theoretically,

Figure 0.1: An example of visual metamerism contingent on point of fixation (the dot in orange) as originally proposed by Freeman & Simoncelli. Both images should be perceptually indistinguishable to each other when fixated at the center. This spatially variant property is an architectural constraint of the human visual system, and is a characteristic that is absent in modern machine vision systems.

understanding the loss of information by matching perceptual distortions that happen in the visual periphery provides us stronger understanding of the computations done by the human visual system. Practically, being able to render metamers in near real-time are critical as a stepping stone for applications with regards to active sensing in machine perception [2], as well as foveated rendering in VR displays. To accomplish this, we use the recent developments in deep learning [3] to develop a new metamer generation model via a foveated style-transfer network [4]. Here, we find that carefully perturbing the encoded image representation in the direction of its texturized-version representation per pooling region, and inverting it – results in a metamer analogous to those proposed earlier by Freeman & Simoncelli.

In the previous two chapters we will have studied the foveated nature of the human visual system by specifically correlating the perception of clutter given a point of fix-

ation with behavioural data such as target detectability, as well as understanding the representations encoded in the visual system that lead to metameric perception of images given texture-driven distortions. Indeed, the human visual system is non-uniform in resolution and this points out to a severe limitation when engaging in visual search, as the observer must make multiple eye-movements as well as scrutinize on a potential location when examining if he/she has found a target. Artificial systems or machines, on the other hand have a set of characteristics that humans (for better or for worse) do not. Machines do not experience fatigue or variable levels of decision-making uncertainty while engaging in visual search. In addition, machines also do not have varying spatial resolution constraints as humans do [5]. In fact, one of the key differences between human and machine vision is that machines have the luxury to 'look everywhere' *at the same time* [6] vs humans [7] who must make search sequentially, and with limitations of visual acuity throughout the visual field. One can imagine that if we gain insight in the decision and search process of a human observer (*e.x.* a TSA agent or a Radiologist) whose search task is to carefully search for suspicious items in an X-ray scan or tumors in a mammogram, then it may be possible to accelerate their search process while preserving their detection performance. Indeed improving the throughput of a TSA agent may significantly reduce security check lines in airports. Analogously, an accelerated version of the same Radiologist may be able to examine more mammograms, while conserving his/her precision, thus enabling faster diagnostics to patients.

Considering the previous limitations and potential benefits to overcoming such limitations, in the final chapter of this thesis we design 2 man-machine collaborative systems that aid humans when engaging in visual search.

The first system we designed was a cognitive optimizer called the *Attention Allocation Aid* (AAAD) which was modelled via previous psychophysical data to estimate how much time and eye-movements are needed to perform visual search as well as taking into account

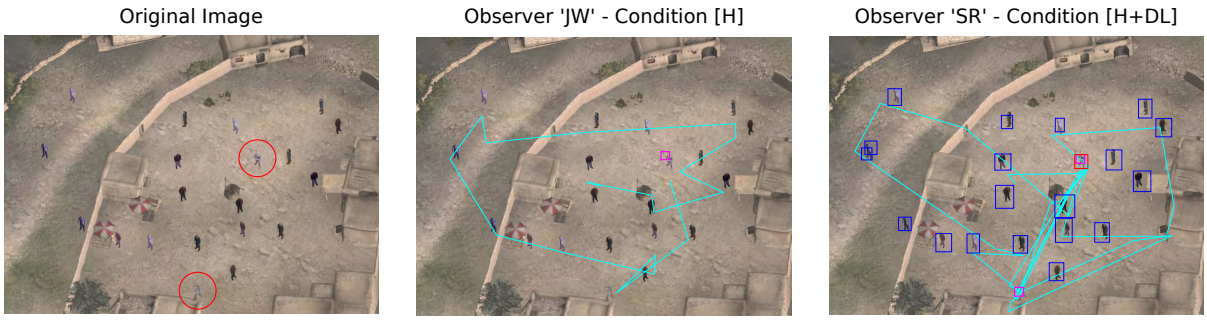


Figure 0.2: An evaluation of potential benefits of a Deep Learning (DL) system aiding in visual search. Left: The original image with targets circled in red. Middle: Boxes in Magenta are clicks that observers did on target location for the [H] condition which is the unassisted human. Right: Boxes in blue represent non-target detections and boxes in red represent target detections of the DL System for the [H+DL] condition where the human performs search with the DL system simultaneously. Middle and Right: Saccadic gaze pattern is plotted in cyan.

the target visibility given the foveated nature of the human visual system and the global levels of clutter in the image [8]. With the goal of trying to understand the elements of visual search satisfaction (knowing when to stop looking), we implement, design and show that human observers increase speed while maintaining performance when engaging in target visual search of the previously mentioned dataset when using the AAAD system.

The second system we designed was a performance optimizer and had the goal of maintaining image throughput and increase performance by supplementing observers with an artificial agent that directly performed visual search. We thus subsequently tried to increase the performance of the human observers with the assistance of an artificial agent, via a widely known object recognition system from the computer vision and deep learning literature: Faster-RCNN [9] as seen in Figure 0.2. To empirically test the hybrid system, we have separate groups of observers engage in two sessions of visual search with and without the Deep Learning (DL) system. We show a set of results that is dependent on the observer's sensitivity compared to the machine in their first session, and more generally that the DL system reduces false alarms across all observers.

Permissions and Attributions

1. The content of chapter 1, with regards to the initial development of a foveated clutter model has been published in the Neural Information Processing Systems (NIPS) 2016 conference. The rest of the chapter which has extended this work with additional experiments is currently in the process of being submitted.

2. The content of chapter 2 regarding the NeuroFovea metamer model has been published at the 2019 International Conference of Learning Representations (ICLR). The work in this chapter was done in collaboration with Aditya Jonnalagadda.

3. The contents of chapter 3 is subdivided in 2 sections, where each section was closely based on a paper. The first section regarding the 'Attention Allocation Aid for Visual Search', has been published at the ACM SIGCHI Special Interest Group on Computer-Human Interaction conference 2017, and was done in collaboration with Jeffrey Peters, Grant Taylor and Amit Surana. The second section was based on the paper 'Assessment of the Faster-RCNN in man-machine collaborative search', which is currently under review at the IEEE Computer Vision and Pattern Recognition (CVPR) 2019 conference, and was done in collaboration with Amit Surana.

Miguel Eckstein was involved in the development of all chapters in this thesis.

Chapter 1

Correlates of Foveated Clutter Models with Perception and Search

1.1 Introduction

An important goal in clutter research has been to develop image based computational models that output quantitative measures which correlate strongly with human perceptual behavior and judgments [10, 11, 12, 13]. Previous studies have accomplished this by creating models that output global or regional metrics to measure clutter perception, and these measures are aimed to predict the influence of clutter on perception via a correlation. However, one important aspect of human visual perception is that it is spatially variant: the fovea processes visual information with high acuity while regions away from the central fovea have access to lower spatial detail. Thus, the influence of clutter on perception can depend on the retinal location of the stimulus and such influences will likely interact with the information content of the stimulus due to crowding [14, 15, 16]. See for example Figure 1.1, where we show intuitively how correlates of target detectability in covert search and the perception of clutter are limited. Indeed, the top-half of the

figure demonstrates how non-foveated models will all output the same score independent of point of fixation for the same image, and this will not correlate well with target detectability. Conversely, the bottom-half of the figure also shows that having a model that directly computes retinal eccentricity only from point of fixation to the target (person) does not account for target detectability given the effects of crowding in the visual periphery that interact with the target. In other words, not all targets are equally detectable even when viewed at the same eccentricity. Having a multiplicative model that uses global clutter and distance from the target is also a limited approach, as there could be images that are globally cluttered, but are not locally cluttered around the target – thus making the target highly detectable. However, if an observer is asked to directly rate the level of clutter of the image, he or she will not consider local levels of clutter as there is no search task imposing the effects of crowding for a target (as there is no target). This problem motivates our work, as we seek to assess the interaction between clutter and retinal location of the image when observers engage in a visual search and clutter judgments task – which are both implicit and explicit measurements of clutter.

To develop a foveated clutter model, we introduce a foveating mechanism based on a peripheral architecture that resembles the biologically inspired log-polar pooling regions [17] that were used in robot vision, and that regained popularity with the metamer model of Freeman and Simoncelli [1] – and stack this architecture onto a non-foveated clutter model to generate a clutter map that simulates the loss of feature response in the visual field due to crowding. The models we will use to compute the perceptual transform that model clutter are a standard collection of models such as Feature Congestion [8, 12], Edge Density [10], Subband Entropy [18] and ProtoObject Segmentation [19, 20] rather than deep end-to-end learned models [21]. We thus produce a new score driven mainly by point of fixation, which implicitly will take into account the effects of crowding in the visual field.

Figure 1.1: An illustration showing the intuition behind foveated clutter models. In general, detecting a target will be easier that further it is in retinal eccentricity from the point of fixation (orange dot). However, the images in the bottom shows an interaction between local levels of clutter around the target and point of fixation given the effects of crowding in the periphery – making some targets harder to detect than others.

This new measure is evaluated on a set of 3 experiments. In the first experiment, we manipulate the eccentricity of a target while observers hold fixation performing covert search to empirically justify the motivation for a foveated clutter model over non-foveated models. However, in the real world humans are making multiple eye-movements when evaluating a scene or looking for an object. Thus in the second and third experiment, we allowed observers to make multiple eye movements as they engaged in a gaze-contingent visual search, and a gaze-contingent clutter judgments paradigm to evaluate how implicit and explicit measurements of clutter change as a function of task, number of eye movements, and points of fixation. Indeed, to our knowledge there has been no systematic evaluation of fixation dependent clutter models across the set of 3 experiments proposed in this chapter: forced fixation search, gaze-contingent search, and gaze-contingent judg-

ments.

The main findings in this chapter are the following: First, we show that the foveated clutter models that account for loss of information in the periphery correlates better with human target detection (hit rate and detectability) across retinal eccentricities than non-foveated models. Second, we design the foveated model to asymptote to the non-foveated model score after observers make multiple fixations in both the search and judgments task. This particular aspect of the design of foveated models enrich the interpretation of clutter and target detectability beyond non-foveated models, as we can compute scores as a function of point of fixation. This is important since observers do not always foveate everywhere in an image: they are sometimes restricted to small areas of search (*e.g.* when one is driving); they encounter gaze-adaptive viewing conditions as in Virtual Reality displays; or they do not have enough time to scan the entire image. Finally, we show under certain assumptions that both the correlations of the foveated models and non-foveated models *vs* the human ratings are similar when observers engage in a judgments task.

In the rest of this chapter we will provide the background of non-foveated models and the motivation for foveated representations, followed by 3 experiments which show the limitations of non-foveated models, and the benefits of foveated models in both visual search and judgment evaluations of clutter.

1.2 Background: Non-Foveated Clutter Models

Regular (non-foveated) clutter models have been defined via computing a perceptual transformation function f of the image stimuli to some perceptual space for the human observer:

$$f : \underbrace{\mathcal{X}}_{image} \rightarrow \underbrace{\mathcal{Y}}_{perception} \tag{1.1}$$

When empirically evaluating clutter models with ground truth, explicit measurements require observers to directly assess clutter, while implicit measurements will measure another behaviour and correlate such performance with clutter. Thus, successful computational models should ideally present strong positive or negative correlations between its score and the following behavioural outputs that may be recorded either explicitly or implicitly:

1. Human clutter judgments such as rankings or ratings (positive correlation): Multiple studies of clutter, correlate their metrics with rankings/ratings of clutter provided by human participants. Ideally, if clutter model A is better than clutter model B, then the correlation of model scores and human rankings/ratings should be higher for model A than for model B [20, 10, 22].

2. Visual search time or response time (positive correlation): Highly cluttered images will require more time for target search, hence more time to arrive to a decision of target present/absent. Under the previous assumption, a high positive correlation value between response time and clutter score are a good sign for a clutter model [12, 23, 22, 24, 11]. One should have in mind that a zero correlation might be an indication of either a bad model design, or a poor experimental paradigm where the task is either too difficult (near zero detectability), or too easy (target pop-out

or ceiling effects).

3. Target detectability (negative correlation): In general, when engaging in target search for a fixed amount of time across all trial conditions, an observer will have a lower hit rate and higher false alarm rate for a highly cluttered image than an uncluttered image [12, 24, 11]. These differences translate into the index of detectability (d') as defined in [25] or Proportion Correct (PC) and clutter score being inversely correlated.

To compute such correlations, each model should output a single one-dimensional score in $\mathbb{R}$, which is lower bounded by zero, and is strictly non-negative (there is no notion of 'negative clutter'). To accomplish this, we usually compute a clutter score Reg (for regular clutter score) which is defined as the composition of a function g which summarizes the multi-dimensional nature of the perceptual transform $f(\circ)$ into a single dimension to make the correlation computable. Thus we have:

$$\text{Reg}(I) = (g \circ f)(I) \tag{1.2}$$

such that $\text{Reg} : \mathbb{R}^D \to \mathbb{R}_+$, f is a perceptual transformation from $\mathbb{R}^D$ to $\mathbb{R}^d$, where d may or may not be equal to D, and g is usually a summary statistic (such as an average, or entropy [12]) which maps the perceptual vector to a single scalar in $\mathbb{R}_+$. It is worth noting that not all models have an intermediate dense representation computed via f as defined in our example. This is only the case of some models such as Feature Congestion and Subband Entropy as we will see in the next subsections. Other models skip this intermediate dense representation and directly compute a score such as Edge Density and Proto-Object Segmentation.

In the next sub-section we will illustrate differences in choice for the function g from where a single clutter score is computed.

1.2.1 Taxonomy of Clutter Metrics

Global Clutter Score: The most popular clutter metric to evaluate models is by computing g over the entire image.

Clutter ROI: Another alternative with regards to computing clutter metrics is local rather than global, restricting g to be computed over a Region of Interest (ROI). This is an approach that has been heavily explored in Asher *et al.* [24] specially when an observer is engaged in visual search and we would like to evaluate how the clutter around the target affects its detectability. Indeed, it could be that a scene is not globally cluttered, but only a collection of regions (one of which contains the target, and is highly cluttered) scores a high region of interest clutter score, but a relatively low global clutter score. The inverse case is also possible: an image may have a high global clutter score and low local clutter score around the target, producing a misleading prediction.

Relative Clutter: An alternative evaluation metric is based on Kullback-Leibler (KL) divergence. KL divergence is a an asymetric distance that computes a dissimilarity between two probability distributions p and q. In previous studies, Deza *et al.* [26] have used KL divergence to approximate how much variation there is between local clutter scores of an equipartitioned 5×4 grid of an image and a uniform distribution. Thus g is computed over the entire image and uses a uniform distribution as a reference. In essence, a highly cluttered image should have a high KL divergence if both distributions are quite different.

1.2.2 Description of Clutter Models

Feature Congestion: Feature Congestion, initially proposed by Rosenholtz *et al.* [8, 12] produces both a pixel-wise clutter score map as a well as a global clutter score for any input image. Each clutter map is computed by combining a Color map in CIELab

space, an orientation map [27], and a local contrast map at multiple scales of a Gaussian Pyramid [28]. One advantage Feature Congestion has over other models is that both the pixel-wise clutter map and global score can be computed in roughly a second. Furthermore, this is one of the few models that can output a specific clutter score for any pixel or Region of Interest (ROI) in an image. Indeed, the function (g_{FC}) is the average of local determinants of the feature covariance matrices of each feature map computed via f. This will be crucial for computational and theoretical reasons that are explored in the Foveated Clutter Models section (Section 1.3).

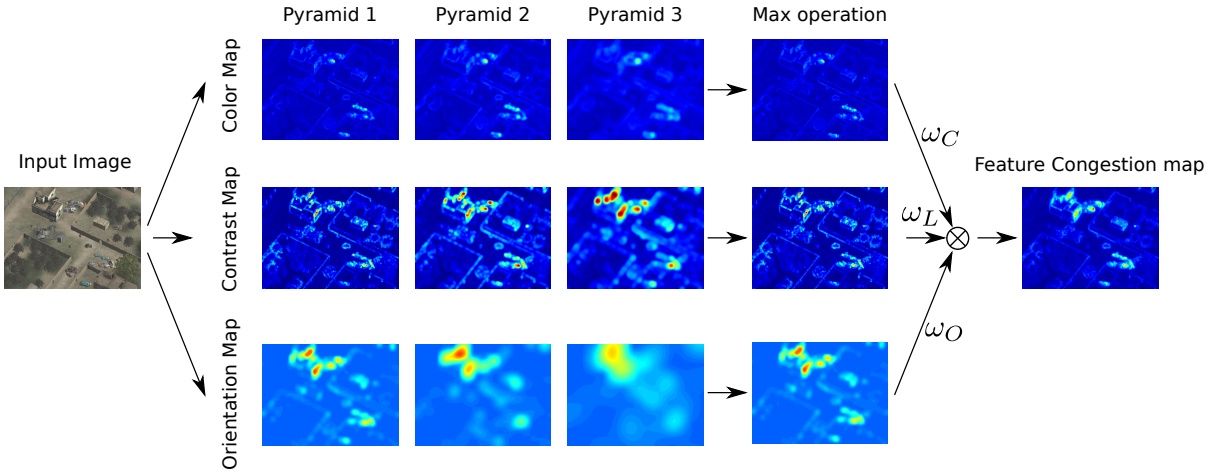


Figure 1.2: The Feature Congestion pipeline as explained in Rosenholtz *et al.* [12]. A color, contrast and orientation feature map for each spatial pyramid is extracted, and the max value of each is computed as the final feature map. The Feature Congestion map is then computed by a weighted sum over each feature map. The Feature Congestion score is the mean value of the map.

Edge Density: The Edge Density metric computes a ratio after applying an edge detector on the input image [10]. We use a Canny filter for our implementation. The final clutter score is the ratio of edges to total number of pixels present in the image. The intuition for this metric is straightforward: the more edges an image has the more cluttered it should be (due to more objects for example). However the model may fail to correlate strongly with search time, for a target that pops out when doing search on

a highly textured scene – as a high density of edges/corner/boundaries in the image will push the ratio to a high value.

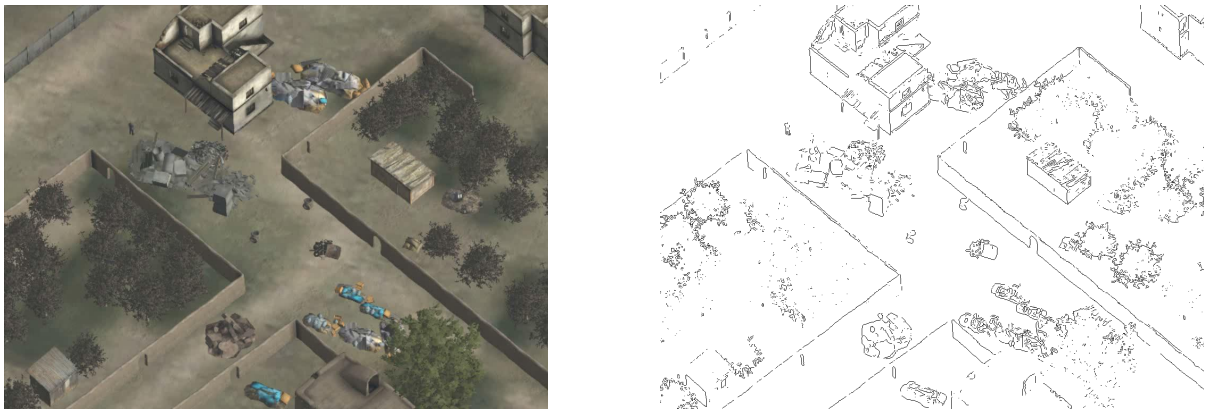


Figure 1.3: The Edge Density pipeline and its intuition: cluttered scenes should generally have more things or stuff and hence more *edges* than non-cluttered scenes. The edge density clutter score is computed via a ratio of all the edge pixels over the total number of pixels in the image.

Subband Entropy: The Subband Entropy metric begins by computing a steerable pyramids [18] decomposition at N scales and K orientations across each channel the input image in CIELab color space. We used $N = 3$, and $K = 4$ for our computations, as implemented in the code provided by [12]. Once each $N \times K$ subband is collected for each channel, the entropy for each oriented pyramid is computed by binning the pixelwise response and they are averaged separately. The 3 responses are finally weighted with a weight of $w_L = 1$ for the luminance channel, and weights of $w_a = w_b = 0.0625$ for the chrominance channels. In this model, the function (g_{SE}) computes the entropy of the steerable pyramid decomposition of the image that serves as an indicator of variability. Thus, the Subband Entropy measures the average entropy over the multi-scale oriented filter responses of an image.

ProtoObject Segmentation: ProtoObject Segmentation proposes an unsupervised metric for clutter scoring [19, 20]. The model begins by converting the image into HSV

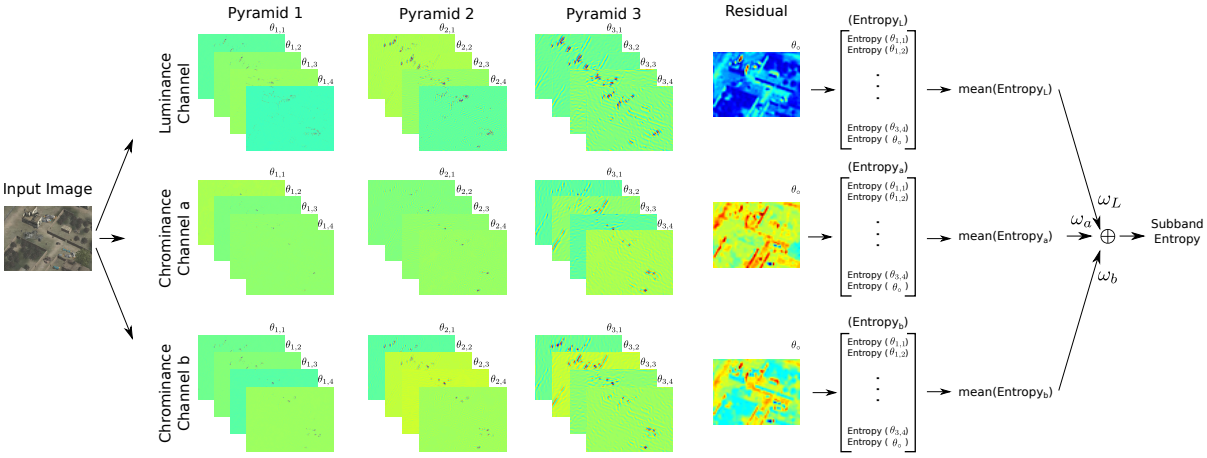


Figure 1.4: Subband Entropy pipeline as explained in Rosenholtz *et al.* [12]. The image is decomposed with a steerable pyramid [18], and the entropy is computed across 3 scales and 4 orientations for each channel in the Lab color space. L is the luminance channel, a is the red-green opponency chrominance channel, and b is the yellow-blue opponency chrominance channel. The average entropy is computed for each channel and a weighted sum of these entropies results in the final subband entropy score.

color space, and then proceeds to segment the image through a superpixel segmentation algorithm [29, 30, 31]. After segmentation, mean-shift [32] is applied on all the cluster (superpixel) medians to calculate the final amount of representative color clusters present in the image. Next, superpixels are merged with one another contingent on them being adjacent, and being assigned to the same mean-shift HSV cluster. Note that the key difference between mean-shift segmentation [33], and ProtoObject Segmentation is the ProtoObject feature vectors do not include $\{x, y\}$ pixel coordinates. We use the SLIC superpixel [31] implementation of VLFEAT [34] with a region size of 40, a regularizer value of 40 for our computations, and a bandwidth of 4 for the meanshift clustering procedure.

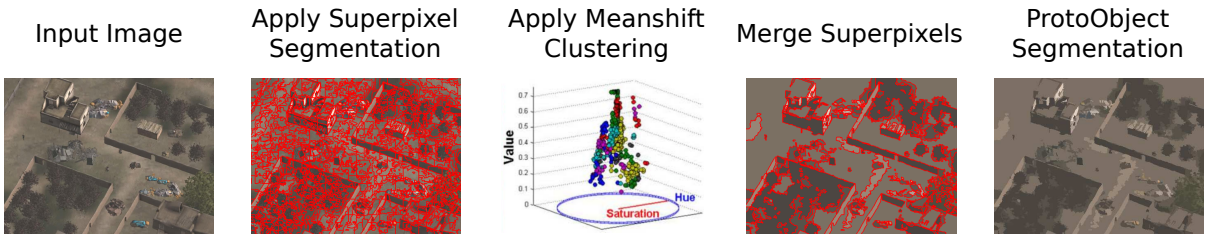


Figure 1.5: The ProtoObject Segmentation as proposed in Yu *et al.* [20]. The input image is segmented via SLIC superpixels, and these regions are clusted in HSV color space via mean-shift. The final score is the ratio of the final number of merged superpixels over the initial number of superpixels, which should provide an indirect estimate of perceptual redundancy.

1.3 Motivation for Foveated Clutter Models

The human field of view can be summarized in two main regions: the fovea and the visual periphery, which are both limited by the photoreceptor density in the retina. In the fovea, there is an exponentially higher density of cones which are tightly packed together within a small $\sim 2\,\text{deg}$ radius around the point of fixation. As we move away from the fovea the cone density decays and the rod density increases evenly tiling the photoreceptor mosaic over a hexagonal lattice – the region within the visual field that lies outside the fovea is called the visual periphery. Thus vision is in reality decomposed into two types: foveal (high acuity) and peripheral (blurred, distorted and low resolution). Both types of visual processing are still active subjects of research in vision science, but the latter still remains quite mysterious [35]. In general, recognition rates of hardly discriminable targets are lower in the periphery than in the fovea. However, some phenomena and targets are easier to detect in the periphery than in the fovea such as motion and a single star in the night sky [36].

A question we should ask ourselves is why should we focus on developing models for foveated clutter perception if although the nature of the human visual system is foveated, we make multiple eye-movements that integrate information spatio-temporally. In other

words, is there a need to decompose the perception of clutter into fixation-like stages? Here we suggest this possibility from a theoretical perspective, given that fixations are the building blocks of eye-movements. Indeed, a reductionist approach might provide theoretical insights of visual clutter perception if we setup a collection of experiments with controlled viewing conditions. In particular we'd like to manipulate the retinal eccentricity of the target from the center of gaze, as well as the number of fixations that observers perform when performing search, or making an explicit judgment.

From a practical perspective, there are some applications where it is not feasible to look everywhere within the field of view, and we would like to estimate the difficulty of detecting a target anywhere in the visual field given a single fixation, or a small number of them. Consider the following examples:

1. A driver can only fixate in a small region of interest in front of him (and only in front of him), and it would be of interest to know if given his field of view how easy or hard it is to detect an unknown obstacle in the path – that may or may not necessarily be a pedestrian [37]. One may argue that that current deep learning systems have near perfect object recognition rates and in near real-time making such foveated models that are dependent on eye-tracking irrelevant [9, 38, 39, 6]. However, the discovery of adversarial examples [40] *a.k.a.* a set of images which are perceptually identical for a human observer, but highly discriminable for the machine may produce fatal accidents in the real world [41] [1]. This would imply that having hybrid-integrated systems that rely on deep learning based object detectors and human real-time estimated gaze data may be a useful solution to the current limitations of machine perception in critical scenarios.

2. A cartographer does not integrate information visually across multiple fixations in

[1] One could say that adversarial images of the same class are 'machine metamers' [1].

the same way for cartographic images [42] and scenes, since they possess different image statistics [43]. Naturally, it is not unreasonable for a human to make sense of a scene potentially within a single fixation [44]. This is not the case of a cartographic image which potentially requires to densely sample and foveate the entire image when looking for a small target. Similarly, radiologists also have their own strategies of visual search [45] where an exhaustive sliding window scan approach would be too inefficient for the number of images they must scan and for their trained eye, as they have intuitions of where and what to look for in mammograms. Having a foveated model of clutter aid observers in visual search by signaling where on the image they might have missed a target provided with a history of fixations may be of interest.

3. State-of-the-art Virtual Reality devices rely on precise human eye-tracking where modeling the amount of peripheral information shown to an observer is critical to preserve functionality and avoid user dizziness. Having a model that can assess how much information is lost in the periphery under gaze-contingent viewing conditions would is relevant. Consider for example a clever design of peripheral widgets such that the information is interpretable, as well as the contrary case where the VR system is purposely trying to fool the human in a game that relies on camouflage via peripheral manipulation.

1.4 Design of a Foveated Clutter Model

Given the limitations of non-foveated models and the before-mentioned motivation for fixation-based models of clutter perception: we propose that the foveated clutter model should output a score which takes into account 3 main terms:

1. A regular model score Reg which acts as a baseline.

2. A Peripheral Integration (PI) coefficient that accounts for the effects of crowding in the periphery which are detrimental for target detection – analogous to the work of Deza & Eckstein [46], and van den Berg *et al.* [22]

3. A bias term that was not previously introduced in Deza & Eckstein [46], given that their approach only included a single fixation study. In addition their approach did not include the assumption that the foveated score should asymptote to the non-foveated score after many fixations.

The first term is global term independent of fixation. The second term will act as a non-linear gain that will modulate the clutter score depending on the amount of crowding around the target given the retinal distance between the target and the point of fixation. The third terms is the bias, which will make the foveated score asymptote to the original non-foveated score after n fixations. With the previous properties mentioned above, we found that the following definition of a foveated clutter score satisfies such conditions:

$$\mathrm{Fov}_I^{p,t} = \mathrm{Reg}_I \times (1 + k\mathrm{PI}^p_{ROI(t)}) \tag{1.3}$$

where $\mathrm{Fov}_I^{p,t}$ is the foveated clutter score expressed as a function of the point of fixation p and the location of the target t, Reg_I is the regular (non-foveated) score of image I, PI is the Peripheral Integration coefficient computed over a Region of Interest (ROI),

and k is a normalization constant that converts the range of the PI, to an interval that is comparable to the output of the original clutter score, such that the foveated clutter score can take into account both terms when computing a foveated clutter score. Indeed, if k is too small the PI will have no effects on the foveated clutter score, but if k is too large, then the Foveated clutter score will be independent of the regular clutter score. We will discuss the computation of k later in this section after we directly evaluate the ranges of the regular clutter scores, and the PI coefficients.

Intuitively, a foveated clutter model that takes into account target search difficulty should score very low when the target is in the fovea (near zero), and very high when the target is in the periphery. Thus, an observer should find a target without difficulty, achieving high detectability rates at the fovea, yet the observer should have a lower target detectability rates in the periphery given local crowding effects around the target. Note that in the periphery, not only should it be harder to correctly identify a target (make a hit), but it is also likely to confuse the target with another object or region affine in shape, size, texture and/or pixel value (false alarms). Under this assumption, we wish to modulate a clutter score by a multiplicative factor called the PI coefficient, where the target and fixation location will implicitly compute the local effects of crowding around the target.

Figure 1.6 provides a run-through of our foveated pipeline when using the Feature Congestion [12] model. The 3 feature maps for color, contrast and orientation are computed across 3 resolutions, and a point of fixation is chosen given actual human psychophysical data, from which the feature maps are then foveated. The foveating mechanism is implemented by performing a max pooling operation over the feature activation maps per each pooling region simulating a winner-takes-all mechanism. The mean pooling operation has also been used van der Berg *et al.*. The same set of averaging coefficients used in the non-foveated maps are used for the foveated maps, and a perceptual difference

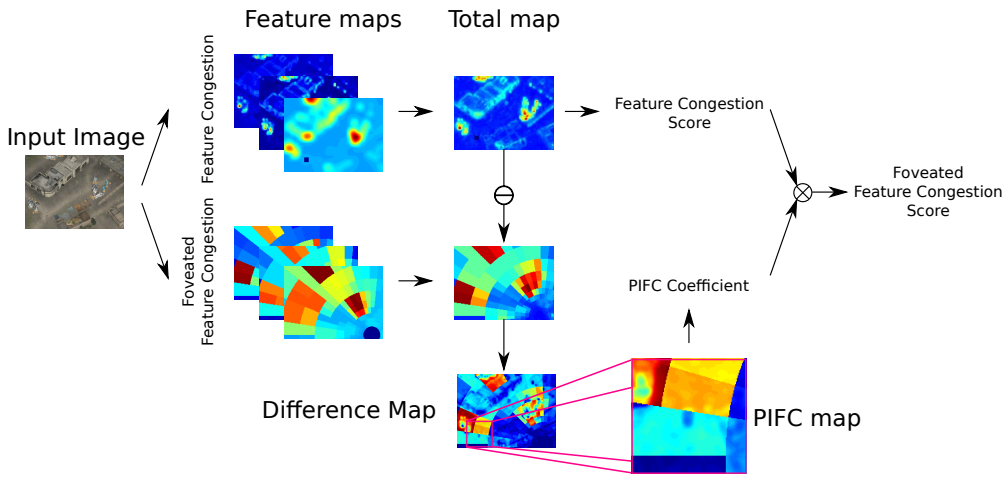


Figure 1.6: Foveated Feature Congestion flow diagram: In this example, the point of fixation is at 15 deg away from the target (bottom right corner of the input image). A Feature Congestion map of the image (top flow), and a Foveated Feature Congestion map (bottom flow) are created. The PI coefficient is computed around an ROI centered at the target (bottom flow; zoomed box). The Feature Congestion score is then multiplied by the PI coefficient, and the Foveated Feature Congestion score is returned. Sample PI's across eccentricities can be seen in the Supplementary Material.

is computed by taking the distance between the regular clutter map and foveated clutter map over a region of interest (ROI). Here, the regular clutter map is defined via:

$$R = f(I) \tag{1.4}$$

and the foveated clutter map, given the point of fixation p, and some foveating function q, is:

$$F^p = q(f(I)) \tag{1.5}$$

The average feature activation loss due to crowding [16] within the ROI is computed via the perceptual distance $D : \mathbb{R}^D \to \mathbb{R}_+$, and we define this value as the Peripheral

Integration (PI) coefficient.

$$\mathrm{PI}_{ROI(t)^p} = D(R, F^p) \tag{1.6}$$

An additional constraint of the model is the selection of the ROI over which the PI is computed. When observers are engaging in visual search, this region of interest over which the PI is computed is a window around the target, and when observers are engaging in a judgments task, the region of interest is the entire image. We will discuss more making this theoretical justification when we experimentally verify these choices in this section, as well as in Experiment 2 (Section 1.7) and Experiment 3 (Section 1.8) respectively. From Equation 1.6 it should also follow that there is little to no loss of feature activation in the fovea given that the maps have very similar (or equal) values, and higher losses of feature activation the further we move away in terms of retinal eccentricity contingent on the local levels of clutter in the visual field. The full details for the computation of a PI coefficient can be seen in Algorithm 1.

The model can also be readily extended to the collection of multiple fixations $\bar{p}$ via the following equation:

$$\mathrm{Fov}_I^{\bar{p},t} = \mathrm{Reg}_I \times (1 + k\mathrm{PI}^{\bar{p}}_{ROI(t)}) \tag{1.7}$$

where

$$\mathrm{PI}^{\bar{p}} = D(R(I), \bigcup_{\bar{p}} F^p(I)) \tag{1.8}$$

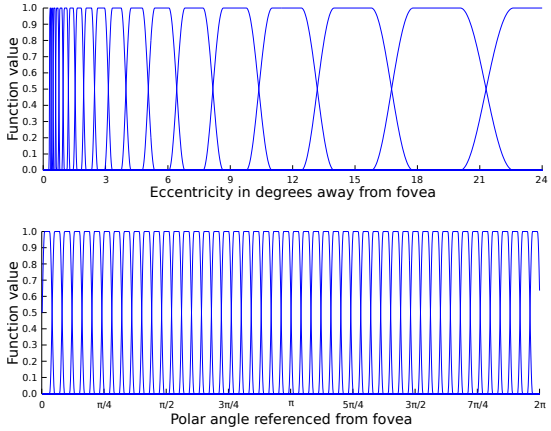
and the ROI over which the PI is computed is an $r \times r$ window when the task is visual search, and is the entire image when the task is judgments given that there is no target (or alternatively, one could argue that the target *is* the image). An example of such task

Algorithm 1 Computation of Peripheral Integration (PI) Coefficient

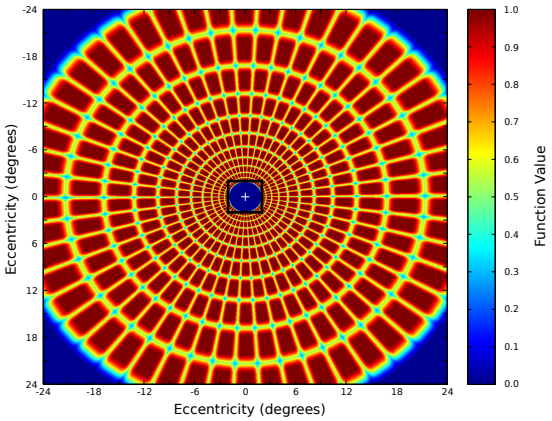
procedure COMPUTE PI OF ROI OF IMAGE I ON FIXATION f
Create a Peripheral Architecture $\mathbf{A} : (N_\theta, N_e)$
Offset image I in Peripheral Architecture by fixation $p : (p_x, p_y)$.
Compute Regular Clutter map R of image I via $R = f(I)$
Set Foveated Clutter $F \subset \mathbb{R}^2_+$ map to zero.
Copy Regular Clutter values in fovea r_0: $F^p(r_0) = (R(r_0))$
    for each pooling region r_i overlapping I, s.t. $1 \leq i \leq N_\theta \times N_e$ **do**
        Get Regular Clutter map (R) values in r_i
        Pool Peripheral Clutter value given Regular Clutter value: $F^p(r_i) = \text{pool}(R(r_i))$
    end for
Crop Foveated map to ROI: $F^p_{ROI} = F^p(ROI)$
Crop Regular map to ROI: $R_{ROI} = R(ROI)$
Choose Distance D between F^p_{ROI} and R_{ROI} map
Compute PI Coefficient $= (D(F^p_{ROI}, R_{ROI}))$
return Coefficient
end procedure

would be providing a clutter rating which is a human judgment. A final characteristic of our model is that if the observer foveates at the target, the PI will asymptote to zero when engaging in search, thus converging to the original clutter score and will on the other hand remain somewhat constant (independent of point of fixation) if the observer is examining the image in a judgments task.

In the rest of this section we will explain two of the elements that are necessary to compute a Peripheral Integration (PI) coefficient. The first is the creation of a human-like peripheral architecture as explained in Section 1.4.1. The second is the choice of the perceptual distance metric between the non-foveated and foveated clutter maps for the PI coefficient as we will discuss in Section 1.4.2. Finally, in the third subsection (Section 1.4.3) we will provide a generalization of the computation of the foveated model to other models that intrinsically do not have an intermediate dense representation f, such as Edge Density, Subband Entropy and Proto-Object Segmentation – thus extending the applicability of the Foveated Clutter model beyond Feature Congestion.



(a) Top: $g_n(e)$ function. Bottom: $h_n(\theta)$ function.



(b) Peripheral Architecture.

Figure 1.7: Construction of a Peripheral Architecture *a la* Freeman & Simoncelli [1] using the functions described in Section 1.4.1 are shown in Fig. 1.7(a). The blue region in the center of Fig. 1.7(b), represents the fovea where all information is preserved. Outer regions (in red), represent the outer pooling regions of the periphery at multiple eccentricities.

1.4.1 Creation of Peripheral Architecture

The Freeman & Simoncelli [1] pooling region model has been tested and used to model V1 and V2 responses in human and non-human primates with high precision for a variety of tasks [1, 47, 48, 49, 50]. It is described by a set of pooling regions that increase in size with retinal eccentricity. Each pooling region is separable with respect to polar angle $h_n(\theta)$ and log eccentricity $g_n(e)$, as described in Eq. 1.10 and Eq. 1.11 respectively. These functions are multiplied for every angle and eccentricity (θ, e) and are plotted in log polar coordinates to create the peripheral architecture as seen in Fig. 1.7.

$$z(x) = \begin{cases} cos^2(\frac{\pi}{2}(\frac{x-(t_0-1)/2}{t_0})); & -(1+t_0)/2 < x \leq (t_0-1)/2 \\ 1; & (t_0-1)/2 < x \leq (1-t_0)/2 \\ -cos^2(\frac{\pi}{2}(\frac{x-(1+t_0)/2}{t_0})) + 1; & (1-t_0)/2 < x \leq (1+t_0)/2 \end{cases} \tag{1.9}$$

$$h_n(\theta) = z\Big(\frac{\theta - (w_\theta n + \frac{w_\theta(1-t_0)}{2})}{w_\theta}\Big); w_\theta = \frac{2\pi}{N_\theta}; n = 0, ..., N_\theta - 1 \tag{1.10}$$

$$g_n(e) = z\Big(\frac{\log(e) - [\log(e_0) + w_e(n+1)]}{w_e}\Big); w_e = \frac{\log(e_r) - \log(e_0)}{N_e}; n = 0, ..., N_e - 1 \tag{1.11}$$

The parameters we used match the rate of growth of the receptive fields of V1 given a scaling factor of $s = 0.25$, a visual radius of $e_r = 24\,\text{deg}$, a fovea of $2\,\text{deg}$, with $e_0 = 0.25\,\text{deg}$ [2], and $t_0 = 1/2$. The scaling factor s (receptive field size/diameter) defines the number of eccentricities N_e, as well as the number of polar pooling regions N_θ from $\langle 0, 2\pi]$. This scaling factor is a free parameter of the model that controls the amount of crowding given the rate of growth of the receptive fields over which the information is pooled. The choice of $s = 0.25$ provides a reasonable estimate over which many of the low level features extracted from the human visual system via some function f such as orientation and contrast are pooled [1]. Throughout this chapter, we will refer to this collection of pooling regions and fovea as a V1 architecture.

1.4.2 Perceptual distance metrics for Peripheral Integration

Let us recall our definition of the Regular Clutter map as R and the Foveated Clutter map as F, both of pixel size S, which are both some computed with some transformation $R = f(I)$, and $F = q(f(I))$, s.t. $q : \mathbb{R}^D \to \mathbb{R}^D$. When integrating information over a region of interest or over the entire image, there are potential metrics that can be used to evaluate how much information has been lost over the selected part of the visual field given the effects of crowding. This loss of information is defined as the Peripheral

[2] We remove regions with a radius smaller than the foveal radius, since there is no pooling in the fovea.

Integration (PI) coefficient as defined in the previous section and is computed with a distance function D, such that:

$$\text{PI} = D(R(I), F^p(I)) \tag{1.12}$$

In this thesis we will focus on 3 distances:

1. **Manhattan Distance** (l_1): Is the average absolute value of the difference between F and R, and is computed via:

$$l_1(R(I), F^p(I)) = \frac{\sum |F^p(I) - R(I)|}{S} \tag{1.13}$$

2. **Euclidean Distance** (l_2): Is the mean square error difference between F and R, and is computed via:

$$l_2(R(I), F^p(I)) = \frac{\sum ||F^p(I) - R(I)||_2^2}{S} \tag{1.14}$$

3. **KL Divergence** (KL): Is the Kullback-Leibler divergence computed between the reference code $\tilde{R}$ and the compressed (foveated) code $\tilde{F}$, via:

$$\text{KL}(\tilde{R}(I), \tilde{F}^p(I)) = \sum \tilde{R}(I) \left(\epsilon + \log \frac{\tilde{R}(I)}{\tilde{F}^p(I) + \epsilon} \right) \tag{1.15}$$

where $\tilde{R}, \tilde{F}^p$ are both the pixel-wise maps transformed into probability density functions.

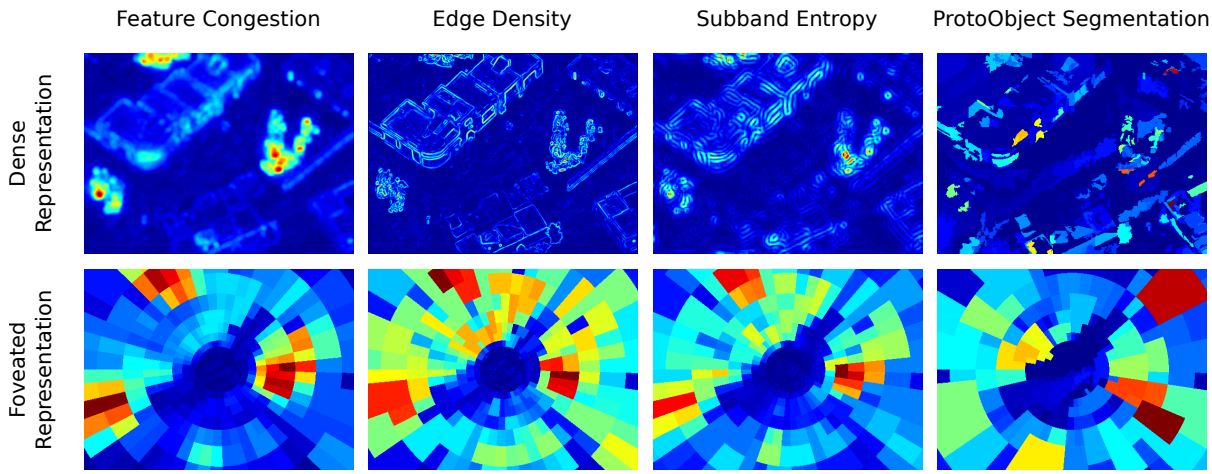


Figure 1.8: Top: Intermediate dense representations via a choice of different functions f for the Feature Congestion, Edge Density, Subband Entropy and Proto-Object Segmentation models. Bottom: Their foveated versions, which are displayed with non-overlapping windows and max pooling.

1.4.3 Generalization of the foveated clutter model

Models such as Edge Density, Subband Entropy and ProtoObject Segmentation have not been designed to produce an intermediate step with a dense clutter pixel-wise representation (unlike Feature Congestion). As we have mentioned earlier in this section, we must find such intermediate step to produce a dense representation over which we can pool the information in the visual field. However, it is hard to find an optimal dense clutter representation without losing the essence of each model. Given that these models do not have a function f, we must resort to computing a proxy of f that still preserves the essence of the model.

For Edge Density, we compute the magnitude of the image gradient after grayscale conversion. For Subband Entropy, we use the 3×4 subbands as dense feature maps, and use the same coefficients to compute a weighted sum over the entropies. In other words, our dense version of Subband Entropy is the steerable pyramid decomposition of the input image. Dense ProtoObject Segmentation was computed by following the

intuition of final number of superpixels over initial number of superpixels, but since this is not applicable at a pixel wise level, we decided to compute multiple ProtoObject Segmentations with different regularizer and superpixel size parameters. These values were: $(0.1, 40), (0.1, 30), (1.0, 40), (1.0, 30), (10.0, 40), (10.0, 40)$ for the regularizer and the region size respectively of the SLIC superpixel algorithm [31] where:

$$c = \frac{\text{regularizer}}{\text{regionSize}} \tag{1.16}$$

and the feature vector over which the clustering is performed is computed for all the pixels given their respective (x, y) coordinate position and intensity $I(x, y)$:

$$\Psi(x, y) = (cx, cy, I(x, y)) \tag{1.17}$$

We later averaged all superpixel segmentation ratio maps – where every map was dense at a superpixel level, and each superpixel score was the initial number of pixels over the final number of initial number of pixels that belong to that superpixel after the meanshift merging stage in HSV color space.

Figure 1.8 (top) shows their corresponding dense representations, and their foveated representations in the bottom. While this step is encouraging given that we can now perform a pooling operation over the visual field, we must now verify that f does in fact preserve the essence of the model. To do so, we must find that composing a summary function g with the intermediate representation function f as shown here:

$$h(I) = (g \circ f)I \tag{1.18}$$

yields a score that is highly correlated with each models original clutter score Reg. A collection of scatter plots per clutter model with the set of all regular clutter scores, and

composed scores via the intermediate representation are displayed in Figure 1.9, where we find that all correlations are statistically significant $p < 0.0001$, indicating an appropriate design of the intermediate representation (choice of f) for the clutter models. The Feature Congestion model has a notable perfect pearson and spearman rank correlation of 1.0 given the nature of the model where: $\text{Reg}(I) = (g \circ f)(I)$ by definition.

1.4.4 Normalization of PI coefficient

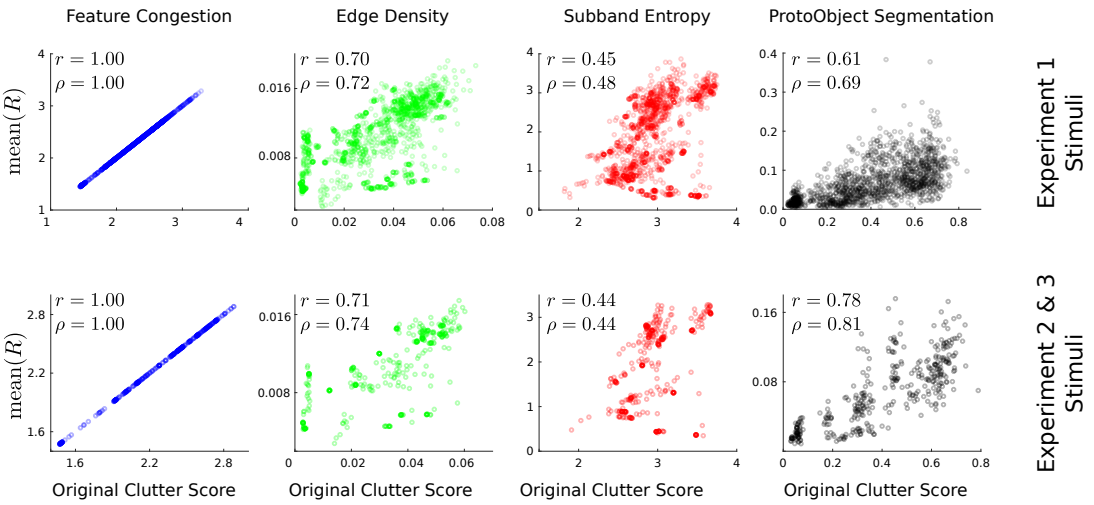


Figure 1.9: A plot of the average dense representation of the function f for each clutter model (mean(R)), plotted against their original clutter score (Reg) for all the stimuli used in the Chapter. All correlations are statistically significant $p < 0.0001$, indicating an appropriate design of the intermediate representation (choice of f) for the clutter models, as well as the motivation for the normalization constant k for the foveated clutter score (Fov).

Though the results of Figure 1.9 are encouraging, we observer that not all models present Reg and mean(R) within the same range. This is the case for Edge Density and ProtoObject Segmentation where these values differ by an order of mangnitude if we compare their x-axis to y-axis *vs* Feature Congestion and Subband Entropy that both have inputs and outputs in the same range. High differences in such outputs may over/under estimate the influence on the PI for the foveated clutter score, motivating

the introduction of the normalization factor k from Equation 1.3. Here, we will find a value of k that renormalizes the range of the PI which is comparable to the original clutter score Reg. As the Regular Clutter score is directly proportional to the mean of the non-foveated clutter map such that

$$\text{Reg} \propto \text{mean}(R) \tag{1.19}$$

where the equality only holds for the Feature Congestion [12] model. Recall that if we define the PI through the l_1 distance, then $\text{PI} = \text{mean}(F-R)$ for some ROI. Consequently, we can define the Foveated Clutter score as directly proportional to the PI coefficient, which yields:

$$\text{Fov} \propto \frac{\text{Reg} \times \text{PI}}{\text{mean}(R)} \tag{1.20}$$

Given that we would like to add a bias such that the foveated clutter score asymptotes to the regular clutter score as the PI goes to zero, we have:

$$\text{Fov} = \text{Reg} + \frac{\text{Reg} \times \text{PI}}{\text{mean}(R)} \tag{1.21}$$

which can be re-expressed as our original definition of foveated clutter score proposed in Equation 1.3, where our choice of k given the previous assumptions will be $k = 1/(\text{mean}(R))$, and we finally have:

$$\text{Fov} = \text{Reg} \times \left(1 + \frac{\text{PI}}{\text{mean}(R)}\right) \tag{1.22}$$

Equivallently, the Foveated Clutter score can also be expressed as a sum if we define

the normalized PI coefficient ($\bar{\text{PI}}$) as:

$$\bar{\text{PI}} = \left(\frac{\text{Reg}}{\text{mean}(R)} \right) \text{PI} \tag{1.23}$$

thus having:

$$\text{Fov} = \text{Reg} + \bar{\text{PI}} \tag{1.24}$$

where

$$\lim_{\bar{\text{PI}} \to 0} \text{Fov} = \text{Reg} \tag{1.25}$$

1.5 Overview of Experiments

The foveated model previously described can be applied to 3 different scenarios that are applicable when performing clutter perception research:

Forced Fixation Search: In the Forced Fixation search, an observer is instructed to covertly detect the presence of the target in the visual periphery as he is restricted to not make any eye movements (hence forcing his/her fixation). In this scenario Eq. 1.3 holds as defined previously with an $ROI(t)$ to be defined over a vicinity around the target such that it encompasses the local effects of clutter around the target.

Visual Search: In the visual search condition an observer is instructed to detect the target after performing a series of eye movements trying to find the target. While the goal of a visual search paradigm is to find the target, such objective may be achieved without the need of foveating the target. In addition, there may be difficult visual search scenarios where an observer foveates at the target and yet still can not detect it. Here, we see the use of the clutter offset term, as an observer makes multiple eye movements, and the foveated clutter model score asymptotes to the non-foveated score.

Clutter Judgments: In the clutter judgments paradigm we are interested in computing what the foveated clutter score should be when an observer is foveating on multiple locations of the image without a search goal in mind. In this condition, we will explore how clutter is independent of point of fixation when the observer must provide a direct assessment of clutter via a score/rating. and observer is not engaged in search.

These 3 scenarios provide the motivation for the experiments that we will perform in the next sections of this chapter.

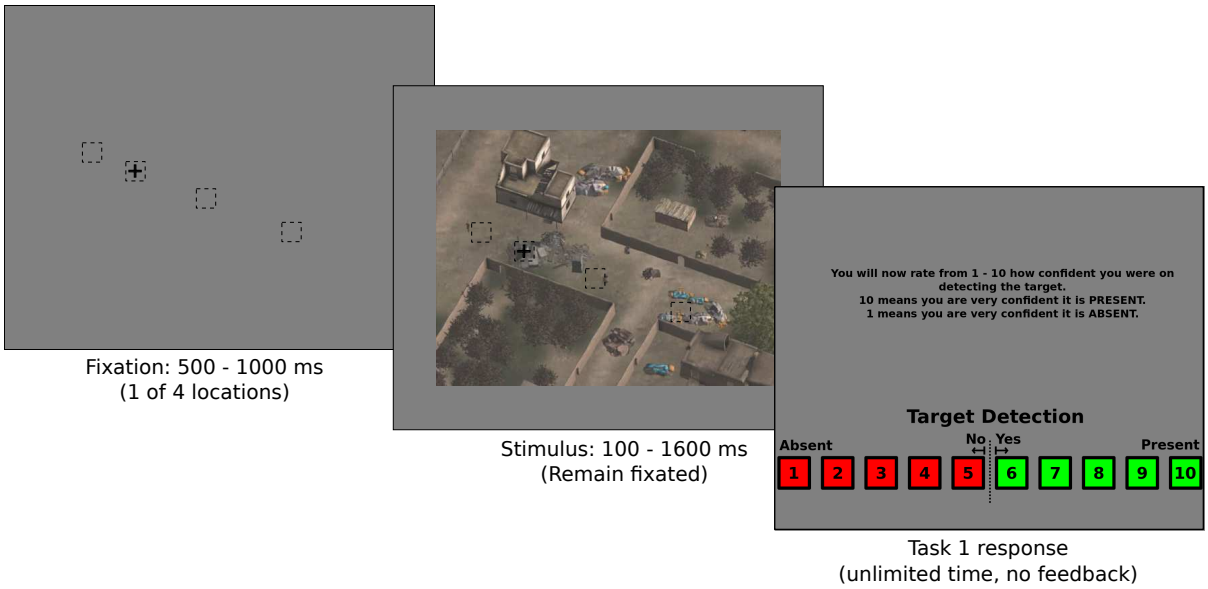


Figure 1.10: Experiment 1: Forced Fixation Search flow diagram. A naive observer begins by fixating the image at a location that is either 1, 4, 9 or 15 deg away from the target (the observer is not aware of the possible eccentricities). After fixating on the image for a variable amount of time (100, 200, 400, 900 or 1600 ms), the observer must make a decision on target present/absent with a 10 point confidence scale.

1.6 Experiment 1: Forced Fixation Search

1.6.1 Methods

A total of 13 subjects, all undergraduates from the University of California, Santa Barbara, participated in a Forced Fixation Search experiment where the goal was to detect a target in the subject's periphery and identify if there was a target (person) present or absent. Participants had variable amounts of time (100, 200, 400, 900, 1600 ms) to view each clip that was presented in a random order at a variable degree of eccentricities that the subjects were not aware of (1 deg, 4 deg, 9 deg, 15 deg). They were then prompted with a Target Detection rating scale where they had to rate from a scale from 1-10 by clicking on a number reporting how confident they were on detecting the target. Participants have unlimited time for making their judgments, and they did not

take more than 10 seconds per decision. The stimuli was half split with an equal number of target present/absent images. There was no response feedback after each trial. Trials were aborted when subjects broke fixation outside of a 1 deg radius around the fixation cross.

Each subject did 12 sessions that consisted of 360 images each of varying levels of zoom (increasing the retinal size of the target and image structures), and different levels of clutter. There were 4 stimuli Sets (each Set consisted of unique images), and each participants viewed each Set 3 times in random order without being aware (4 sets $\times$ 3 times = 12 sessions). Every set also presented the images with aerial viewpoints from different vantage points (Example: Set 1 had the target at 12 o'clock, while Set 2 had the target at 3 o'clock). To control for any fixational biases, all subjects had a unique fixation point for every trial for the same eccentricity values. All images were rendered with variable levels of clutter. Each session took about an hour to complete. The target was of size $0.5\,\text{deg} \times 0.5\,\text{deg}$, $1\,\text{deg} \times 1\,\text{deg}$, $1.5\,\text{deg} \times 1.5\,\text{deg}$, depending on zoom level of the stimuli, and the varying levels of image complexity consisted of different terrain layouts and structures in the scene.

The Hit Rate for each image was computed by gathering the aggregate hits across all 13 observers for every image seen at a different eccentricity ($1\,\text{deg}, 4\,\text{deg}, 9\,\text{deg}, 15\,\text{deg}$). Hits were considered trials where observers assigned a score of 6 or higher for stimuli that had a target present. For each image, the same trials were combined for each one of the 12 sets, and for each eccentricity, since each of the sets represented either the same image or a rotated version of itself. Notice that a rotated version of each scene might not have the same clutter score (though they are very similar), despite that these values are approximate versions of each other. These scores are averaged for the computation of the clutter score of the 4 rotated versions.

1.6.2 Parameter selection for the Peripheral Integration coefficient

There are many factors to take into account when computing the PI coefficient, including: if we should include the target or not in the computation of the clutter score (as the target itself could potentially contribute to clutter – see Asher *et al.* [24]), the choice of the perceptual distance function (l_1, l_2 or KL divergence), the pooling operation: max vs mean pooling, as well as the size of the Region of Interest (ROI). Indeed, this amounts to $2 \times 3 \times 2 \times r$ ways of computing the PI, where r is the number of discrete steps over which we can define the ROI window to be. We restricted our experiments to ROI windows of 4×4, 6×6, 8×8, 10×10, 12×12 and 14×14 degrees of visual angle (d.v.a) centered around the target.

To find an appropriate choice of parameters for the PI computation, we decided to plot the Pearson and Spearman rank correlation between the PI coefficient and target detectability (d') for all the possible settings which can be shown in Figure 1.11. We will discuss these findings in the next subsections.

Computation of Target detectability (d')

One might be tempted to use the strict definition of d' on a per eccentricity basis:

$$d'_z = \Phi^{-1}(\mathrm{HR}_z) - \Phi^{-1}(\mathrm{FA}_z) \tag{1.26}$$

where z is one of the 4 retinal eccentricity where the target lies on, Φ is the cumulative distribution function of the normal distribution, HR is the target hit rate, and FA is the target false alarm rate.

However a problem arises when trying to compute the false alarm rate, as the false alarm rate per eccentricity is *undefined*. Indeed, when the target is not present the target

could be at *any* of the 4 retinal eccentricities, so it is unreasonable to assign a specific false alarm rate for a retinal eccentricity, and they should rather all be computed equally as if the observer would respond with a false alarm equally for *any* retinal eccentricity for a specific image.

As a solution we will be computing a d' per eccentricity, which is a partially pooled d'_z (only over false alarms) which is computed via the following equation:

$$d'_z = \Phi^{-1}(HR_z) - \Phi^{-1}(FA_{\bar{z}}) \tag{1.27}$$

where $\bar{z}$ is the collection of all 4 eccentricities. The hit and false alarm rates are computed by combining the trial decisions over the collection of 13 observers, and these are computed per each image resulting in a vector of 4 (d') detectability scores per image. In the following sub-sections, we will focus our evaluations on the 100ms trials, as they provide stronger differences in detectability and avoid ceiling effects.

Choice of Target inclusion or removal

In our implementation, including the target required computing the default PI, while removing the target required padding the target with NaN's in the pre-pooling stage for all the models. In the study of Asher *et al.* [24], target removal was implemented by padding the target location with a black patch. We did not perform such target removal strategy as changing the image information might directly affect the clutter score in an unexpected way (rather than by directly removing those pixels).

We find that the choice of whether to include or remove the target in the computation for the PI coefficient is not relevant for our particular experimental setup. This can be verified visually by observing the similarity of the plots of Target Included and Removed in Figure 1.11 in a column-wise manner. This conditions holds true across all clutter

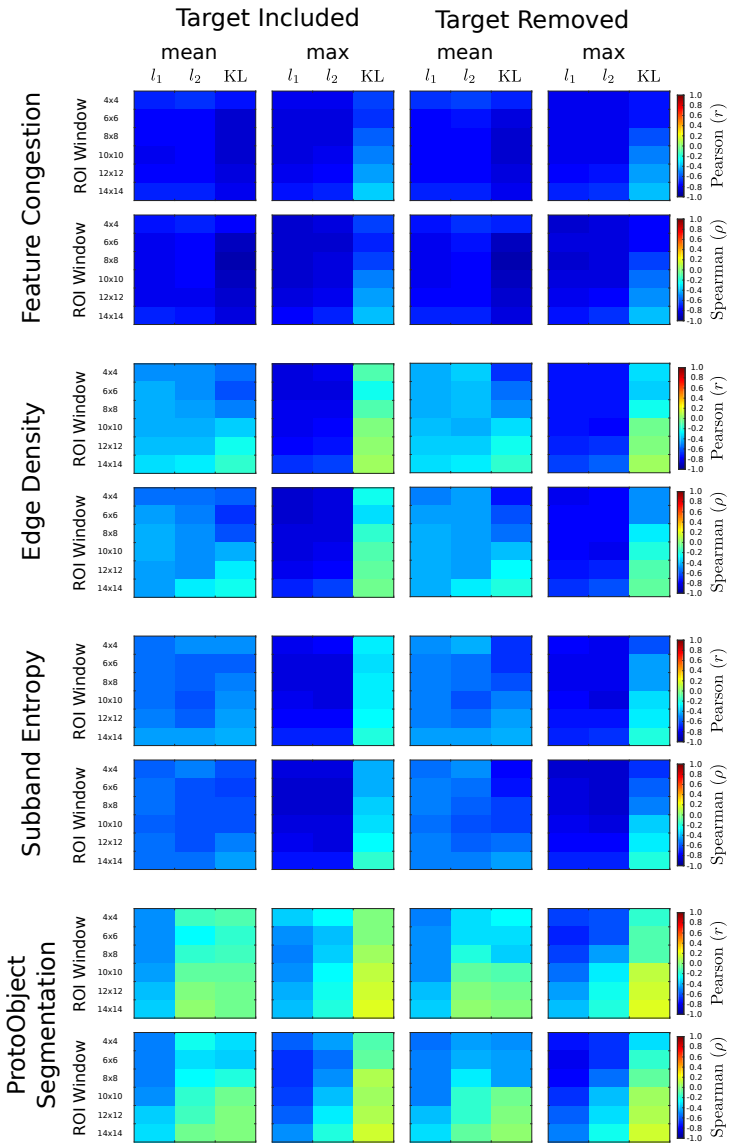


Figure 1.11: Empirical Evaluations of the Peripheral Integration (PI) coefficient in terms of their correlations with target detectability (d') across multiple conditions for each clutter model. We evaluated target present/absent in the computation of the PI, pooling operation, perceptual metric D, and choice of ROI window size around the target.

models, and across all perceptual distances. The differences become more obvious when one must choose the ROI over which to make the computations, as a smaller ROI will by default include a stronger effects of the target. The figure shows such small differences which are most notable at the 4×4 ROI (recall that the target is roughly 0.5×0.5 for

the images used to plot the figure).

In a visual search task of finding an arrow in a map, Rosenholtz *et al.* [12] found similar results when they correlated log reaction time with clutter scores for conditions of target present and absent, yet they only computed a single clutter score for the image which was performed before the target was superimposed on the background. One might find this result quite surprising, as we would think that the appearance of a target influences its likelihood of detection compared to the background image, as well as its local vicinity. Indeed, in the previous study [12], target present or absent did influence reaction times when engaging in visual search. We believe one of the reasons why we've found that the correlations between d' and PI do not change much when including the target in the computation is that the target generally maintains the *same appearance* throughout the entire experiment. In other words, we do not have different small targets that vary heavily in color, or contrast. This is perhaps one of the current limitations of our model – which is that it directly does not compute any property of target appearance. To our knowledge, no current clutter model that is evaluated with target detectability includes target appearance as factor of the model. This aspect and other limitations will be discussed more in detail in the General Discussion section.

Thus, as we are not directly modeling target appearance, in the rest of this chapter we will compute the PI by removing the target from the image.

Choice of Pooling Operation

There is a general trend across all models that the pooling operation that produces the strongest correlations is the max pooling operation. In addition, the choice of pooling operation is heavily coupled with the choice of perceptual distance. In general proper metrics such as l_1 and l_2, produce higher values of PI when the difference between the regular and foveated map are accentuated. The same is not the case for the mean

response, although there is biological evidence that supports averaging of features per receptive field when studying crowding upon oriented gabors in the periphery as studied in Parkes *et al.* [51]. Indeed, though the max pooling operator and the l_1 perceptual distance produces the strongest correlations across most models, another alternative is the mean pooling, with KL divergence as perceptual distance – in defense of averaging models that account for crowding [16]. However, winner takes all mechanisms (the max pooling model) have also been shown to account for the effects of distractors upon orientation acuity as seen in Palmer [52], Palmer *et al.* [53], and also explored in Morgan *et al.* [54]. Moreover, both the mean and max pooling operations are variations of the generalized form of Minkowski summation as referred to in Morgan *et al.* [54], and Graham [55]:

$$R = (\sum R_i^{\beta})^{1/\beta} \tag{1.28}$$

where R is the total stimulus response, and R_i is the individual response of stimulus i. One can see that when $\beta = 1$ we have the averaging model, and when $\beta = \infty$ we have the max (winner takes all) model.

In the rest of this chapter the operation we will choose to compute the PI is max pooling, though we do not discard mean pooling – coupled with KL divergence as the perceptual distance – as an excellent alternative.

Choice of Perceptual Distance

There is little to no change in correlation of PI *vs* target detectability between the l_1 and l_2 metrics for the choice of perceptual distance as shown in the Figure 1.11.

However, there is an interaction between the type of pooling operation and the distance metric of KL Divergence. Indeed, the correlations are stronger when the pooling operation is the mean rather than the max for KL divergence. It is not clear why this

might be the case, although one possibility might be that since KL divergence requires each map to be transformed into a 2D probability density function, the difference in values diminishes post re-normalization from the max value computed in the map. The l_1 difference might not be as strong as when each map is re-normalized with an averaging component, while re-scaling both non-foveated and foveated clutter maps after the mean operation accentuates their differences when the log ratio is computed between the Regular and Foveated maps for the KL divergence. While future work should investigate such reasons, we have found that indeed the similar choice of mean pooling and KL divergence was used in the Crowding Model of van der Berg *et al.* [22], that produces a score similar to the PI coefficient computed over the entire image.

In the rest of this chapter we will use the l_1 distance as it produces similar correlations to the l_2 metrics, and both of these produces stronger correlations that KL divergence for any choice or region of interest and pooling operation. In addition, the l_1 distance outputs a smaller range of values than the l_2 distance, which reduces the range of the scores produced by the foveated model. Again, we'd like to re-iterate that the coupling of mean pooling and KL divergence is a plausbile alternative that requires thorough investigation.

Choice of size of ROI window

We found that the highest correlations for target detectability for the l_1 metric and max pooling operation, target absent condition, were achieved when computing the PI over a 6×6 and 8×8 ROI window around the target . This region of interest agrees with the size of the ROI found in Asher *et al.*, when computing local clutter scores around the target of similar size to ours. Asher *et al.* [24], linked this result to the span of effective search size of 5 deg as shown in Bertera & Rayner [56]. However this is a finding that is not directly applicable to our experiment as observers are not engaging in visual search

with eye movements, but rather covert search under forced fixation.

It may be possible that higher areas of the ventral stream compute such difference operation over fixed regions in the cortical surface. Potentially, pooling region based mechanisms of crowding occur at the early stages of visual processing, but the total region over which further areas compute such crowding effects are fixed. This might be due to difference in convergence rates of photoreceptor to LGN, LGN to V1 neurons, V1 to V2 neurons, and so forth. This rate of convergence at some point may stabilize, and requires thorough investigation of linking this result to the classical diagrams of Felleman & Van Essen [57], which could explain a fixed ROI that produces high correlations with target detectability after pooling. It is also worth mentioning that the average receptive field size of a neuron (from macaque) in inferior temporal area TEO is 5.8 deg (as compiled in Kravitz *et al.* [58]). This may suggest that IT may play a role in clutter perception, more specifically the integration of loss of information given post-crowding effects in the visual field. It has also been suggested by Eriksen & James [59] that the attentional system may act as a *'zoom lens'* independent of our center of gaze. Thus it could be that the fixed window of 6 deg is related to an attentional window over which information is processed. Future experiments should potentially investigate the role of attention on our foveated clutter model by having cross-hair like cues on the outer part of the image, as well as investigating the neurophysiological ([60]) or cognitive basis for the 6 deg integration window.

In the rest of this chapter, we will use a 6 deg window for our the analysis in experiments.

1.6.3 Analysis of PI coefficient, Foveated and Regular clutter score

We used 11 images from our analysis in the 100ms viewing time condition which are all shown in Figure 1.12. These images all varied in levels of image complexity, but had the target positioned at different locations (retinal eccentricities) that interacted differently with the scene background.

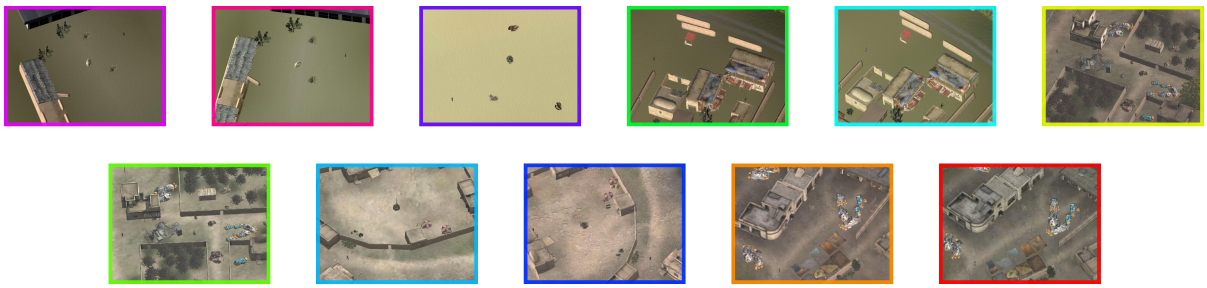


Figure 1.12: The sub-collection of the images that were analyzed in the 100ms condition where we varied the retinal eccentricity of the fixation to the target at values of 1, 4, 9, 15 deg. The colors bounding the image have the same assignment as those plotted in Figure 1.13, Figure 1.14, and Figure 1.15.

We computed the target detectability as an aggregate from all the observers given Eq. 1.26, for each image, and these scores are shown color coded in Figure 1.13 (top). Here, one can see the limitation of regular (non-foveated) clutter models: each image has a single clutter score that is in a wide range of the target detectability axis – leading to limited interpretability of the clutter model if one is to assess how easy it is to detect a target given its clutter score. This may be one of the reasons why many experiments as performed in Rosenholtz *et al.* [8, 12], Asher *et al.* [24], van der Berg *et al.* [22], which have a visual search component, require the target to be placed far from the center when search begins at a center fixation. If a user designer or engineer would want to determine how likely an observer is going to find a target given a non-foveated clutter score, the predictions would give a broad range in d', limiting its usefulness specially if the observer

of interest is to restrict his gaze to a certain region or a single fixation.

However, the results of correlating target detectability with the PI coefficient which takes into account such losses of visual information contingent on local levels of clutter around the target, are more interpretable and give smaller ranges of d'. These are shown in Figure 1.14 (top), along with the change in PI as a function of retinal eccentricity per each image (bottom). For example for Feature Congestion PI, we can confidently say that *independent* of the image, if its PI score is 3, then the d' will be approximately between $[0, 1]$ (very low detectability). Conversely if the PI score is below 1, then d' is approximately between $[1, 3]$. These types of inferences of target detectability as a function of the regular clutter score can not be made, as the detectability for our images spans through all values of the clutter scores. At the end of this section, we will indeed verify the advantages of a foveated *vs* non-foveated model when such correlations are done per eccentricity.

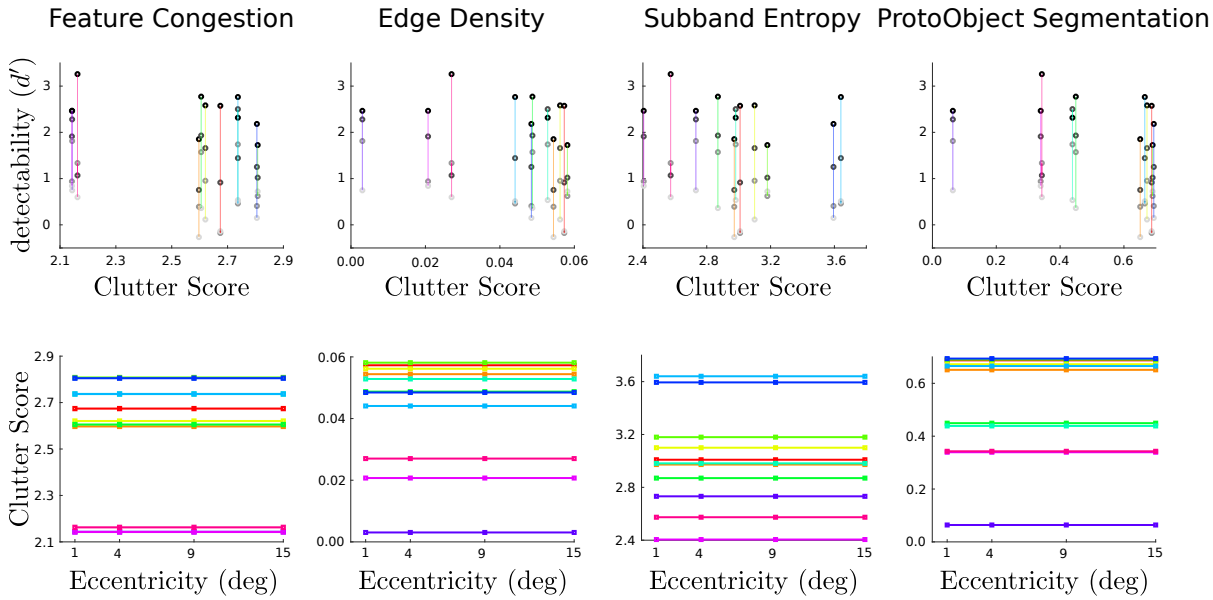


Figure 1.13: The limitations of regular (non-foveated) clutter models under forced fixation tasks. The clutter score does not change as a function of target eccentricity limiting the relationship of clutter and target detectability.

Tables 1.1 report both Pearson correlations (r, linearity) and Spearman rank correlations (ρ, monotonicity) as evaluations for all models with respect to d' – including the Foveated Clutter scores as seen in Figure 1.15.

Pearson Correlation (r) between detectability (d') and Clutter Model				
Model	Feature Congestion	Edge Density	Subband Entropy	ProtoObject Segmentation
Reg	-0.21	-0.25	-0.18	-0.33
Fov	-0.78	-0.72	-0.78	-0.77
PI	-0.82	-0.75	-0.80	-0.72
Spearman Rank Correlation (ρ) between detectability (d') and Clutter Model				
Model	Feature Congestion	Edge Density	Subband Entropy	ProtoObject Segmentation
Reg	-0.16	-0.25	-0.20	-0.31
Fov	-0.76	-0.71	-0.79	-0.81
PI	-0.83	-0.77	-0.84	-0.79

Table 1.1: Pearson r (top) and Spearman Rank ρ (bottom) correlations between models and d'. The Fov and PI correlations are all statisically significant with $p < 0.0001$, while no regular clutter model correlation except for ProtoObject Segmentation is statistically significant ($p < 0.05$).

In Deza & Eckstein [46], we reported the correlations between hit rate (rather than detectability) and found the following scores: For Feature Congestion: $r_{FC} = -0.19$, $r_{FC+FoV} = -0.82$; Edge Density: $r_{ED} = -0.21$, $r_{ED+FoV} = -0.76$; Subband Entropy: $r_{SE} = -0.19$, $r_{SE+FoV} = -0.77$; ProtoObject Segmentation: $r_{PS} = -0.30$, $r_{PS+FoV} = -0.74$ – all of these coherent with our updates results in term of d'. Analogous to our previous results, we found that the highest Pearson correlation for non-foveated models was achieved by ProtoObject Segmentation, but the highest correlation for foveated models was achieved by Feature Congestion, even after the addition of the normalization constant $k = 1/(\text{mean}(R))$ and the inclusion of the bias term (Reg). In general, these results also show that both PI coefficients alone, and foveated models had higher correlations of target detectability and target hit rate rather than non-foveated models across eccentricities.

Our model is also different from the van der Berg *et al.* [22] model since our peripheral

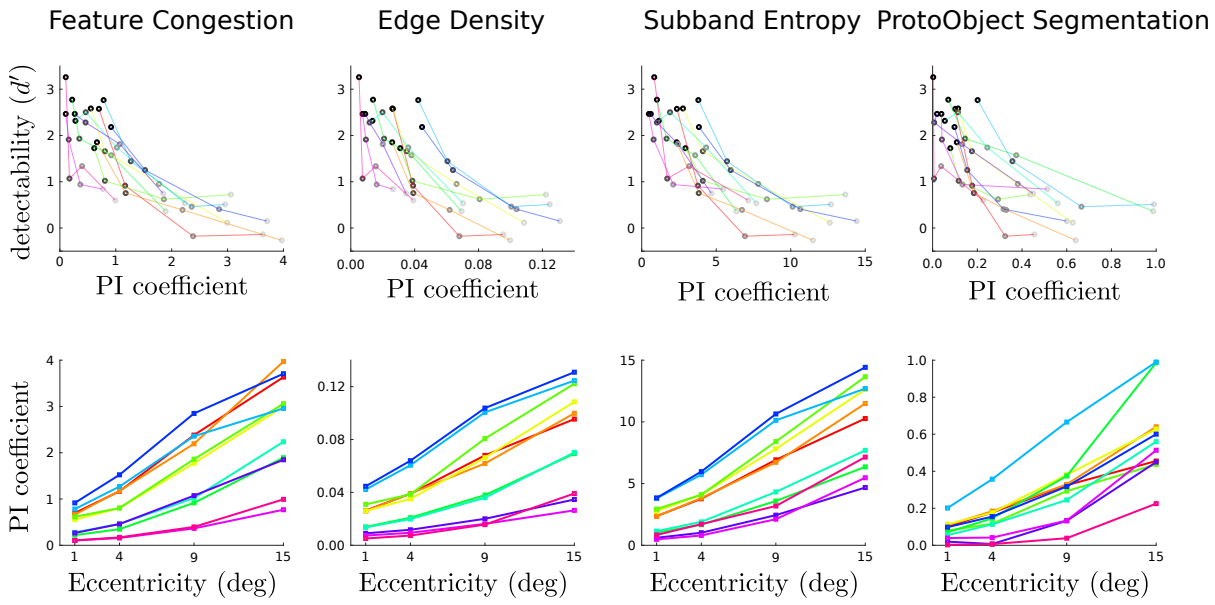


Figure 1.14: The role of the Peripheral Integration (PI) coefficient for each clutter model *vs* target detectability (d') (top) and across multiple eccentricities (1, 4, 9, 15 deg) (bottom) for the 100ms condition. Each color represents a different image from the stimuli across both plots. In the top plot, the eccentricities are coded in discrete ascending gray scale tone from black (1 deg) to white (15 deg).

architecture uses: a biologically inspired peripheral architecture with log polar regions that provide anisotropic pooling [61] rather than isotropic gaussian pooling as a linear function of eccentricity [22, 62]; we used region-based max pooling for each final feature map instead of pixel-based mean pooling (gaussians) per each scale – which allows for stronger differences as seen in Figure 1.11; this final difference also makes our model computationally more efficient running at 700ms per image, *vs* 180s per image for the Crowding model ($\times$250 speed up). A home-brewed Crowding Model applied to our forced fixation experiment resulted in a correlation of $(r(44) = -0.23 \pm 0.13, p = 0.0469)$, equivalent to using any of the non-foveated models. Perhaps the reason for the Crowding Model to not perform so well in our images is that the window over which we are integrating information is defined around the target, rather than around the entire image. In addition, van der Berg *et al.* [22] evaluated there model with clutter rankings and reaction times on

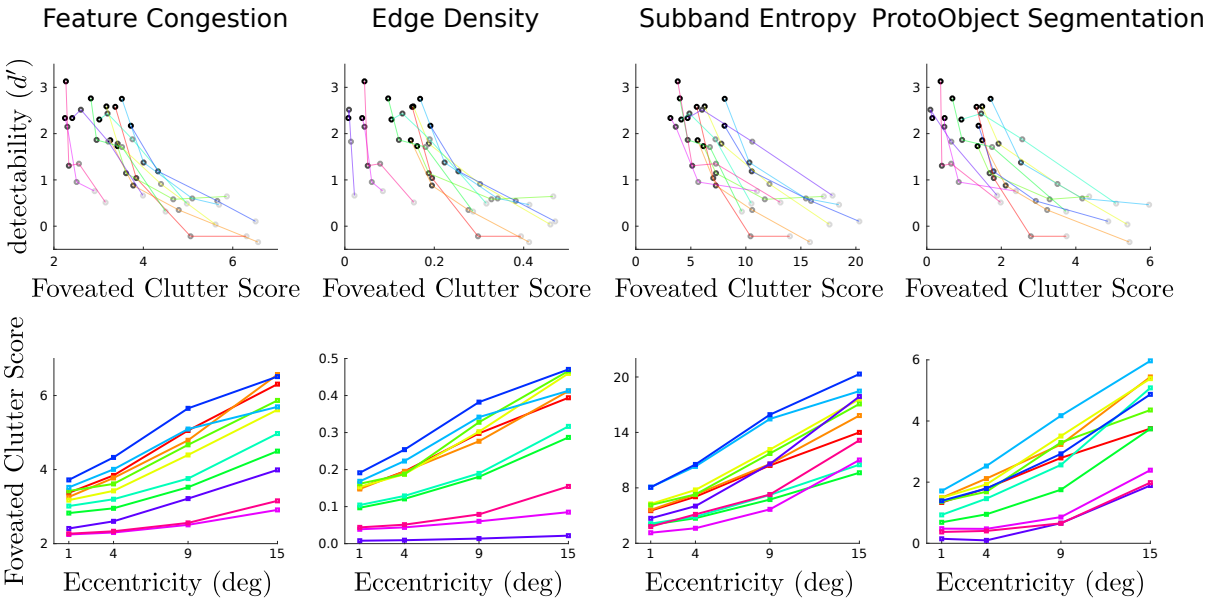
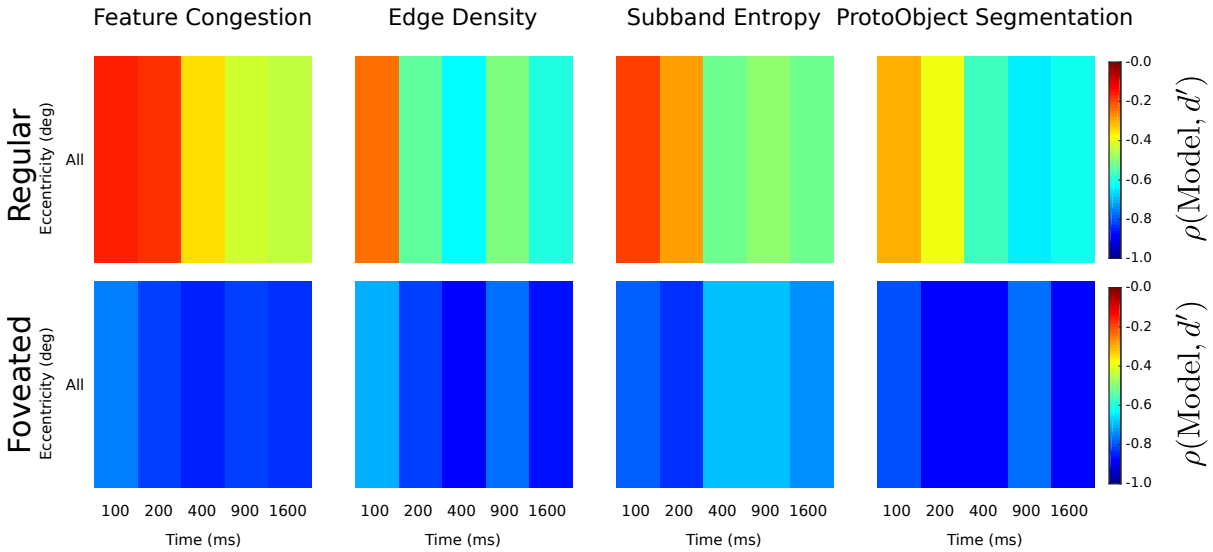


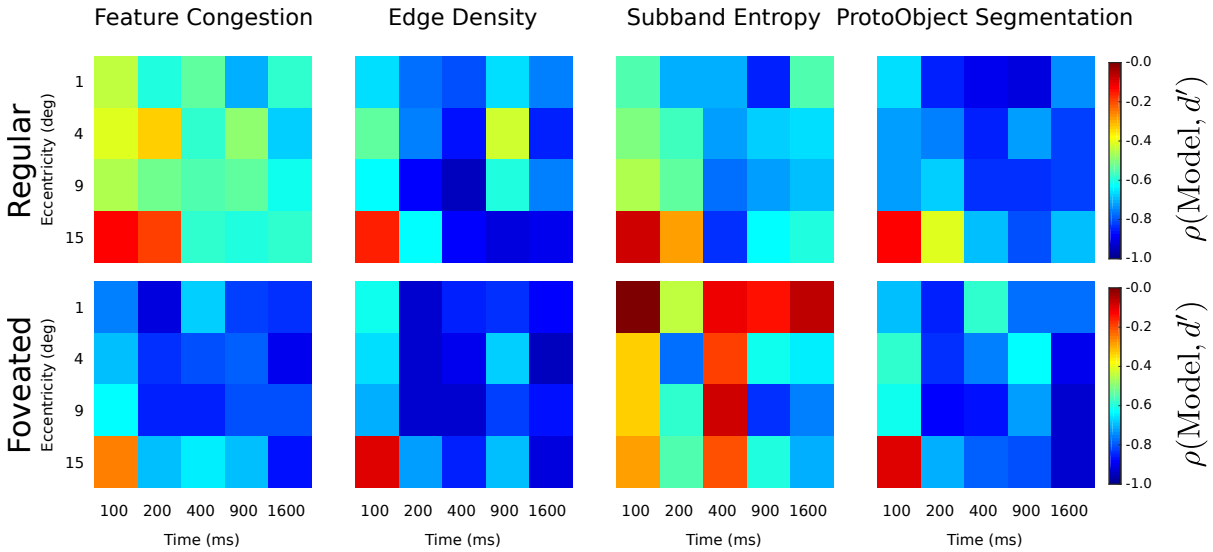
Figure 1.15: The Foveated model overcomes the limitations of the regular model, by integrating the PI coefficient as a factor in its computation. The model produces a shape similar to the PI coefficient, but in a range that is not strictly lower bounded by zero, but rather by the global clutter score of the image – a useful property when considering multiple fixation scenarios as we will see in Experiment 2 and 3.

images that possess different statistics to ours such as the maps of Rosenholtz *et al.* [12], and the collection of objects in a gray tone background in Bravo & Farid [23] respectively. Thus, differences in stimuli might also affect model performance as certain models may required different hyper-parameter tuning to achieve better performance contingent on the visual stimulus.

Finally, in Figure 1.16 we answer the initial question that was motivated at the beginning of the chapter: How does the model generalize across multiple viewing conditions? The figure shows how not only does the foveated representation outperform the non-foveated regular model across all viewing times, for all models, but that it also shows increments in terms of Spearman Rank correlations at the *per eccentricity* level. Correlation increments might be stronger for the time condition (where retinal eccentricities are collapsed) vs the eccentricity partition as each eccentricity has only 12 images vs 48



(a) Evaluating the Foveated and Regular clutter models per viewing time.



(b) Evaluating the Foveated and Regular clutter models per viewing time and eccentricity.

Figure 1.16: The two subplot show how the correlations of the clutter models and target detectability d' vary as a function of viewing time and/or retinal eccentricity. We see strong improvements for collapsed eccentricities (top), and an interaction per eccentricity: mild improvement for $(1, 4\,\text{deg})$, high improvement for $(9, 15\,\text{deg})$.

in the collapsed condition per viewing time [3]. Indeed, there is an interaction as these

[3]The images tested for the 100ms condition only were composed of 11 as we had to remove one image that was mistakenly labeled with incorrect eccentricities (though monotonically increasing), though including the image in the analysis does not affect our results.

differences are accentuated at 9 deg and 15 deg for all viewing times. Consequently, a foveated clutter score that is driven by a peripheral representation enriches our understanding of clutter through implicit correlations with target detectability across a wide variety of viewing conditions.

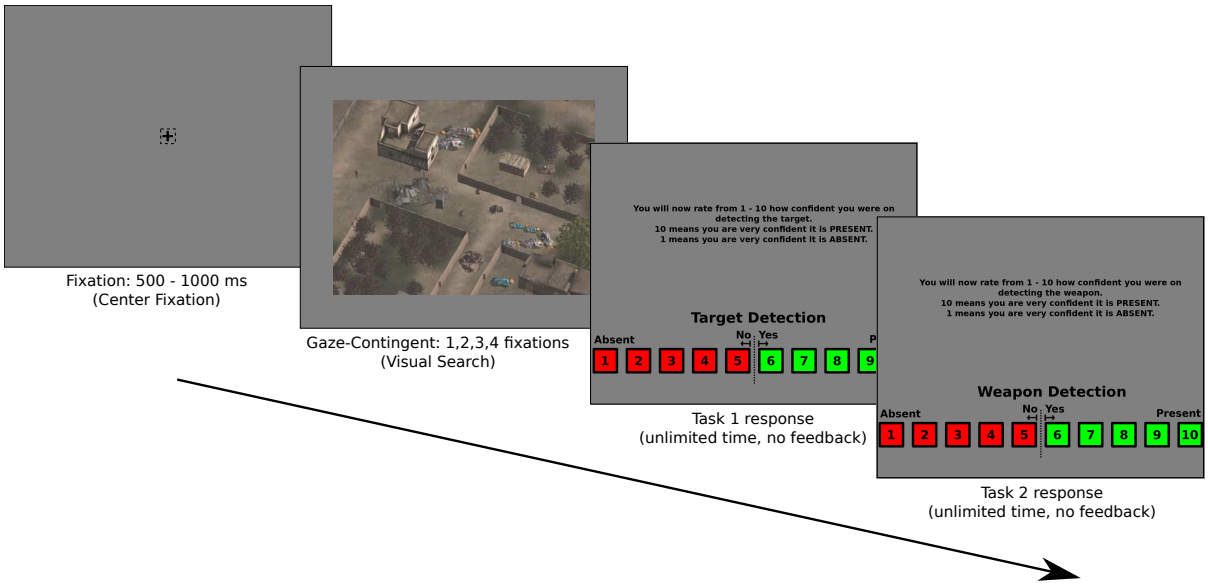


Figure 1.17: Experiment 2: Gaze-contingent visual search. Each trial begins with a fixation cross at the center of the screen. The observer presses the spacebar and the stimulus appears for a limited amount of fixations ending after the *n-th* saccade, although observers believe the experiment is randomly halted by time. Two response screens later appear requiring the observer to input his confidence rating on target (person) and weapon detection, both of which appear 50% of the time contingent on trial and person present respectively. There is no feedback at the end of each trial.

1.7 Experiment 2: Gaze-Contingent Visual Search

In Experiment 1 we tested the use of the foveated clutter model and the Peripheral Integration (PI) coefficient for the simplest case of visual search: forced fixation search. In this experiment, we extend our analysis to a more plausible scenario of visual search where observers are allowed to make eye movements as done in the real world. However in our setup, observers must not only determine if the person is present or absent in each image, but also specify if the person is holding a weapon. We call both of these tasks: target detection and weapon detection which should possess varying levels of difficulty for the same image stimuli.

As our foveated model is driven by fixation location, we designed an experimental

setup where the eye-tracker is polling observer gaze in real-time, thus terminating search after n saccades, restricting the observer to only process n fixations to make a decision. Extending the model to the multi-fixation scenario was discussed in Section 1.3, where we preserve the region of interest (ROI) at $6 \times 6\,\text{deg}$ around the potential target. Integrating information over multiple fixations is achieved via the equation also mentioned in Section 1.4, as computing the PI for more than one fixation is non-trivial:

$$\text{Fov}_I^{\bar{p},t} = \text{Reg}_I \times (1 + k\text{PI}^{\bar{p}}_{ROI(t)}) \tag{1.29}$$

where

$$\text{PI}^{\bar{p}} = D(R(I), \bigcup_{\bar{p}} F^p(I)) \tag{1.30}$$

and $\bigcup$ is the min operation performed for the cumulative foveated representation map $F^p(I)$ of all fixations that each observer must perform when viewing an image. Figure 1.18, illustrates how we extend the foveated model for multiple fixations showing in detail how the Region of Interest is selected, and how the foveated single-fixation information maps are integrated forcing the area within the ROI to asymptote to the original dense clutter score. We will experimentally verify that this enables the (unnormalized/normalized) PI coefficient to go to zero as the number of fixations increase, furthermore pushing the foveated clutter score to the regular clutter score.

Thus in this sub-section, we will show how the foveated model generalizes from the single fixation to the multi-fixation scenario.

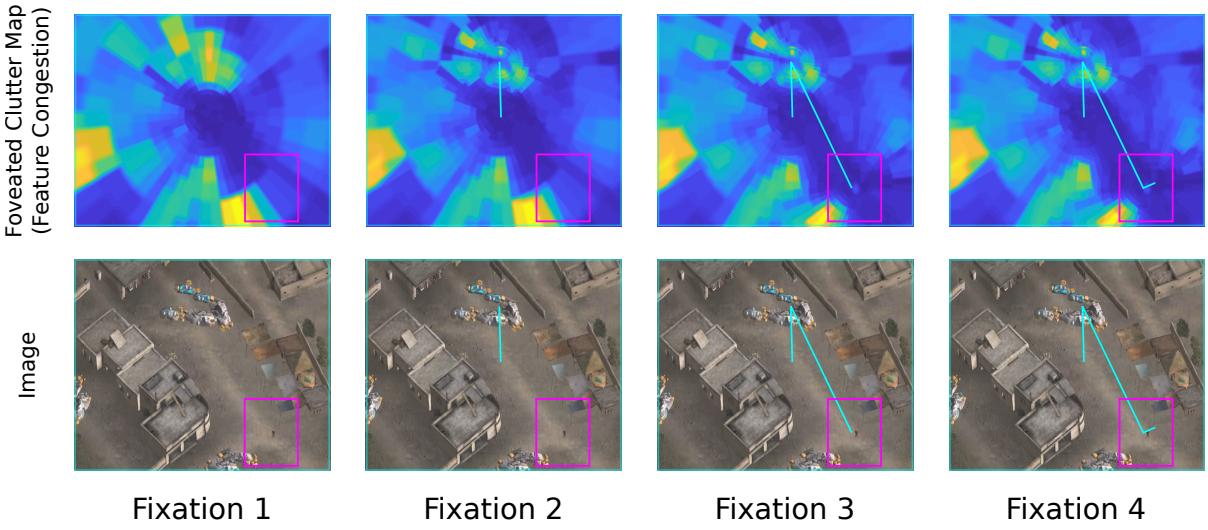


Figure 1.18: An illustration that shows how the PI is computed over multiple fixations and its asymptotic nature (to zero) as observers fixate closer to the target. The ROI is displayed through the pink 6×6 deg box which has been superimposed on the sub-figures for visualization purposes. A real sample saccadic trajectory is superimposed in cyan.

1.7.1 Methods

A total of 6 human observers participated in a visual search task where the goal was to find a person (target) in the scene and determine if the person was holding a weapon or not. Each observer did 4 sessions, where each session consisted of 8 blocks/sets of 80 trials each, totaling 2560 trials. The observer started each trial by forcing center fixation on a cross for a short amount of time (half a second) until an image was displayed. The image stimuli was displayed at 1024×760 pixels resolution at 0.022 deg /px, equivalent to 22.5×16.7 d.v.a. However, the image stimuli seemed to be shown for a variable amount of time to participants, yet it was displayed contingent on the number of saccades that the observer made. In other words, observers did not know that the trials were governed through a gaze-contingent setup, as they were told that the images would be displayed for a variable amount of time. Deceiving the observers as we debriefed them with the instructions was necessary to *not* have them plan their saccades ahead of time. The

number of potential saccades made per trial was evenly distributed between $\{1, 2, 3, 4\}$, and they were randomly assigned in a way such that each image was viewed with a different number of fixations. After the image was shown, two response screens were displayed one after the other. Each response screen had a 10-point rating scale asking the observer how confident he/she was on detecting the person in the scene as well as the weapon.

Participants did not receive any feedback at the end of each trial or throughout the experiment. However, at the very first session, observers saw 30 randomly selected scenes from the stimuli with either target present or absent for 2 seconds each, to get familiar with the type of stimuli that they would encounter during the experiment as well as the appearance and potential locations of the target and weapon. Figure 1.17 shows a diagram of our experimental setup.

We'd like to emphasize that the collection of image that we used for this Experiment 2 (and Experiment 3) were a subset of those from the total collection of images that were viewed in Experiment 1, where we only selected those with small targets to reduce potential ceiling effects of target detectability.

1.7.2 Computation of fixation location for the Gaze-Contingent paradigm

We subsampled the gaze data into fixations given the saccadic thresholds computed by eye movement speed over $22\,\mathrm{deg}/s$, and acceleration over $4000\,\mathrm{deg}/s^2$. As the starting point of the (n)-*th* saccade is not always the same as the ending point of the $(n-1)$-*th* saccade, given the known effects of ocular drift, the (n)-*th* fixation was computed through their average. For the 1st fixation we used the starting saccade position, and the last fixation position was the end of saccade $(n-1)$. The SR EyeLink 1000 terminated the

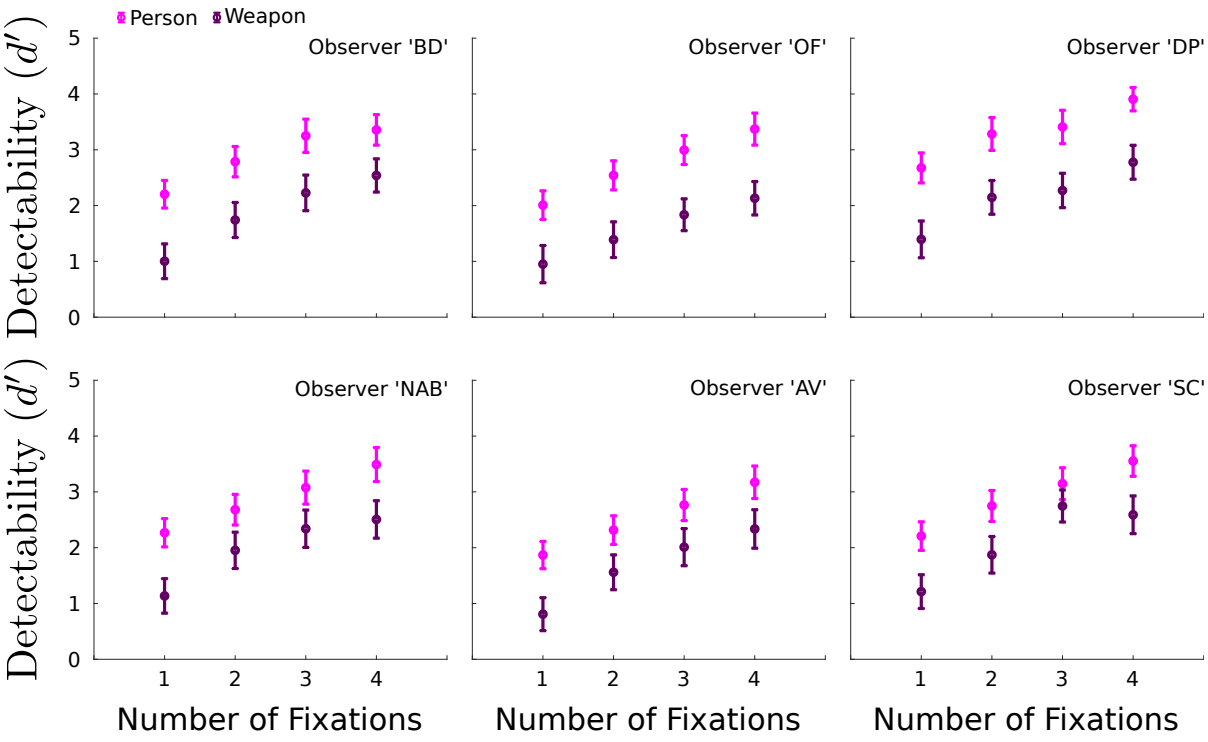


Figure 1.19: Performance of 6 observers when they engage in visual search in a gaze contingent experiment where we terminate the trial after the *n-th* saccade. There are 2 main results: 1) Observers perform better at person detection (pink) than weapon detection (purple), and; 2) performance in d' increases as a function of number of eye movements. Error bars denote the 68% confidence interval after bootstrap sampling.

trial at the exact end of the (n)-*th* saccade in real-time, thus restricting the processing of information from the landing fixation.

1.7.3 Analysis of Person and Weapon Detectability

Target detectability was computed following the un-modified definition of d', as we do not need to pool over multiple eccentricities (or other conditions for the false alarm rate which we were limited by in Experiment 1), given that: 1) we have identical images for target present and absent; 2) the computation of d' only requires us to compute hit rate and false alarm rate for the same trials per image and number of fixations. Hits for

both target and weapon present were considered as those images where observers input a score of at least 6 or above on target present stimuli. Performing this simplification is reasonable given that our response screen had two color codes (red & green) for the target absent/present response. We used the same criteria to report false alarms for target absent stimuli.

The *per* observer hit rate and false alarm rate was computed via bootstrapping (10000 samples) the aggregate of the trials across all images that were viewed with the same number of eye movements and with the target present/absent condition respectively, which totaled 640 trials. The number of trials over which the bootstrap was performed for the weapon detection was 320, given that half of the image stimuli had no target present – and weapon present/absent trials are only valid if the target (person) is present.

The behavioural results of our gaze-contingent experiment are summarized in Figure 1.19, from which we can draw two conclusions: The first is that person detectability is significantly higher than weapon detectability across all fixations. This is naturally the case given that determining if a person is holding a weapon requires fine-grained discrimination, usually to the point at which observer must *fixate* at the target. In consequence, this requires observers to scrutinize the target with a potential extra fixation around it. Target detection on the other hand, can still be done reasonably well above chance in the periphery (See Experiment 1 in Section 1.6). The second result is that detectability increases as a function of number of fixations – a known result in the visual search literature, that was also performed in Najemnik & Geisler [63] for finding gabor patches embedded in spectral $(1/f)$ noise. This should be the case as observers gain more information as they continually explore the image increasing their hit rate and decreasing their false alarm rate. These changes in performance translate into an increase in d'.

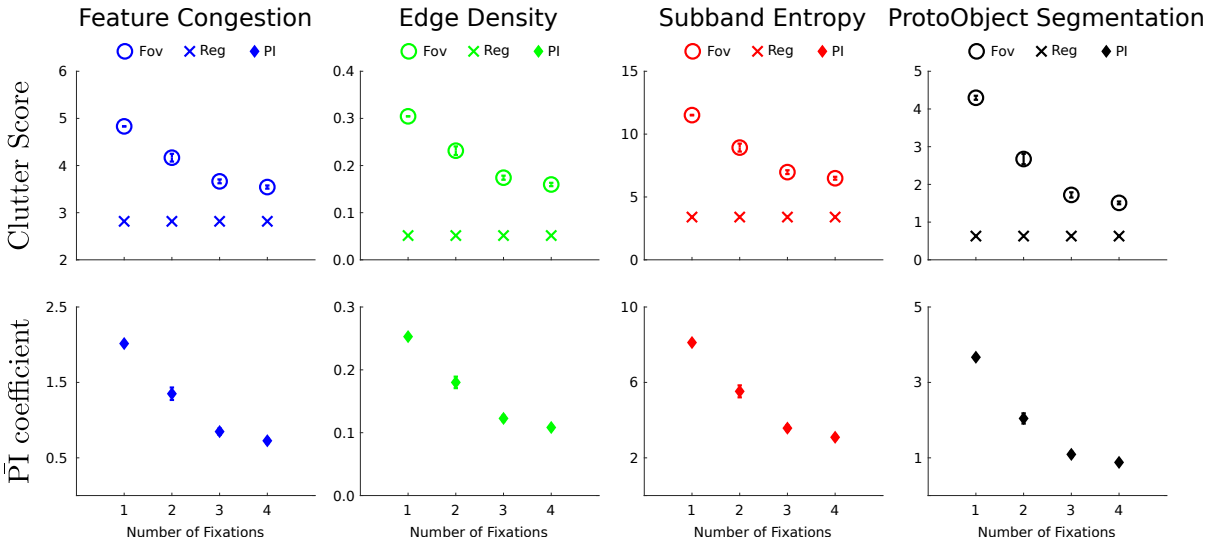


Figure 1.20: A collection of plots where we averaged scores from valid trials across the 6 observers showing the Foveated and Regular scores (top), as well as the normalized PI coefficient (bottom). As the normalized PI asymptotes to zero – proportional to an increase in number of fixations, and inversely proportional to increase in d' (Figure 1.19), the Foveated Clutter score asymptotes to the Regular clutter score. Error bars show the standard error of the mean.

1.7.4 Analysis of PI coefficient, Foveated and Regular clutter score

An early criteria we proposed in Section 1.3 with regards to the motivation of a Foveated Clutter score, is that as the number of fixations increases, the Foveated Clutter Score (Fov), should asymptote to the Regular Clutter score (Reg). Recall from Equation 1.24, that we have the re-written version of the Foveated Clutter score as the addition of the Regular Clutter score and the normalized Peripheral Integration ($\bar{\text{PI}}$) coefficient:

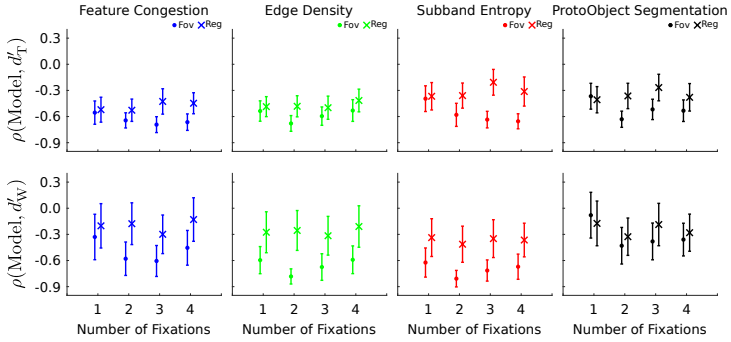
$$\text{Fov} = \text{Reg} + \bar{\text{PI}} \tag{1.31}$$

where the PI decreases as the number of eye movements increases. This criteria is satisfied across all models (Feature Congestion, Edge Density, Subband Entropy and

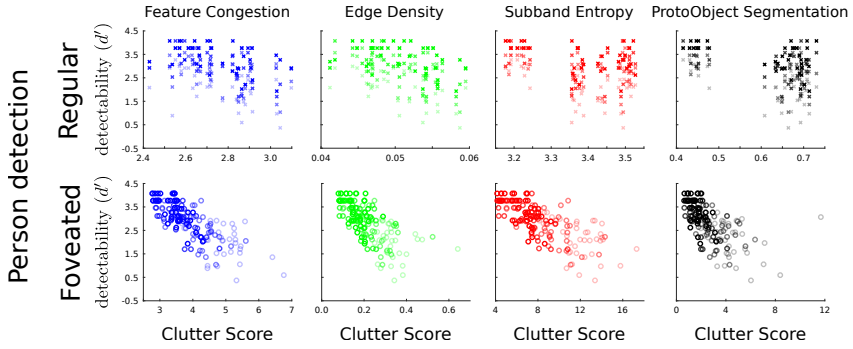
ProtoObject Segmentation) as shown in Figure 1.20 (top), where the Foveated Clutter score asymptotes to the Regular Clutter score. We also separately plotted the normalized PI coefficient ($\bar{\text{PI}}$) in the bottom of the figure, where we see how on average the clutter score of all images is affected given the decrease of ($\bar{\text{PI}}$). Indeed, this plot represents the aggregate across all clutter scores of the image stimuli of Experiment 2. At the *per image* level we find that the scores decay faster or slower depending on the difficulty of search and the observers eye movements.

To appropriately analyze the benefits of using a foveated clutter model over a non-foveated (regular) model we computed the Spearman rank correlation (ρ) coefficients between the aggregate observer detectability for target (d'_T) and weapon (d'_W) against the clutter model score per each image. It should be the case that the foveated clutter model yields stronger negative correlations between d' and clutter scores across all fixations as well as per fixation. Indeed, we previously explored in Section 1.6, how the foveated model has stronger correlations across and per eccentricities. Similarly, we would like to show that as eye-movements increase, the foveated model still presents a representational advantage over non-foveated models.

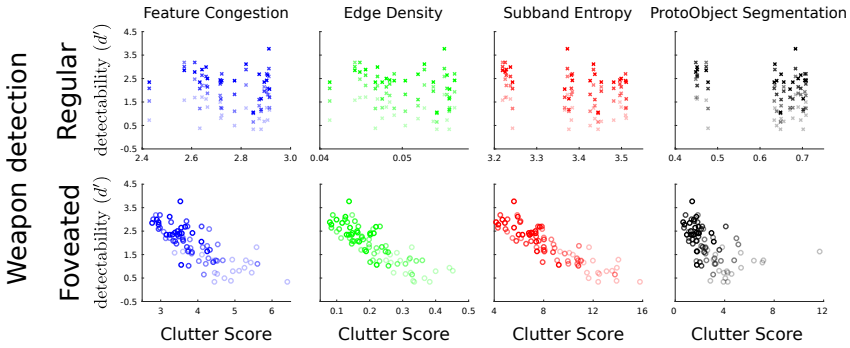
Figure 1.21(a), shows such effects where a foveated representation of clutter has stronger negative correlations across all fixations for all models. Here we show that the foveated model still overperforms that non-foveated model for both the target (person) detection task and the weapon detection task *per fixation*, thus proving that the foveated representation is robust to more than one type of visual search scenario. Indeed, this is an encouraging result that extends those presented in Experiment 1 when analyzing correlates of the different models and their representations at the *per* eccentricity level. In addition Figure 1.21, shows the collection of images and their aggregate d' scores that were used to compute the correlations of Figure 1.21(a). In general, the tendency we see is similar to those of the previous Section, where the image clutter scores are transformed



(a) Spearman Rank Correlations of model and detectability (d')



(b) Target/Person Detection



(c) Weapon Detection

Figure 1.21: Top: The Spearman rank correlation coefficient (ρ) for both target/person (top) and weapon (bottom) detection across all models for both the foveated and non-foveated (regular) representations as we increase number of fixations. Error bars show the 68% confidence interval after bootstrapping. Middle and Bottom: A collection of scatter plots where each image is paired with its detectability (d') score as a function of eye movements. Each fixation is color coded given the alpha intensity of the data point from darkest (alpha value = 1.0) for the first fixation, to lighest (alpha value = 0.25) for the 4th fixation.

to another clutter space (the foveated/peripheral representation) such that all images are re-arranged given the Peripheral Integration (PI) coefficient such that they lie very close to a line. This effect is more noticeable for Weapon detection than for target detection as not restricting eye movements increases ceiling effects for target detectability, contrary to our forced fixation experiment. Indeed, this last result is perhaps the most encouraging one from our set of experiments, as the Foveated Clutter score enriches our understanding of clutter by providing a general framework for clutter perception through the implicit measure of target detectability across multiple fixations.

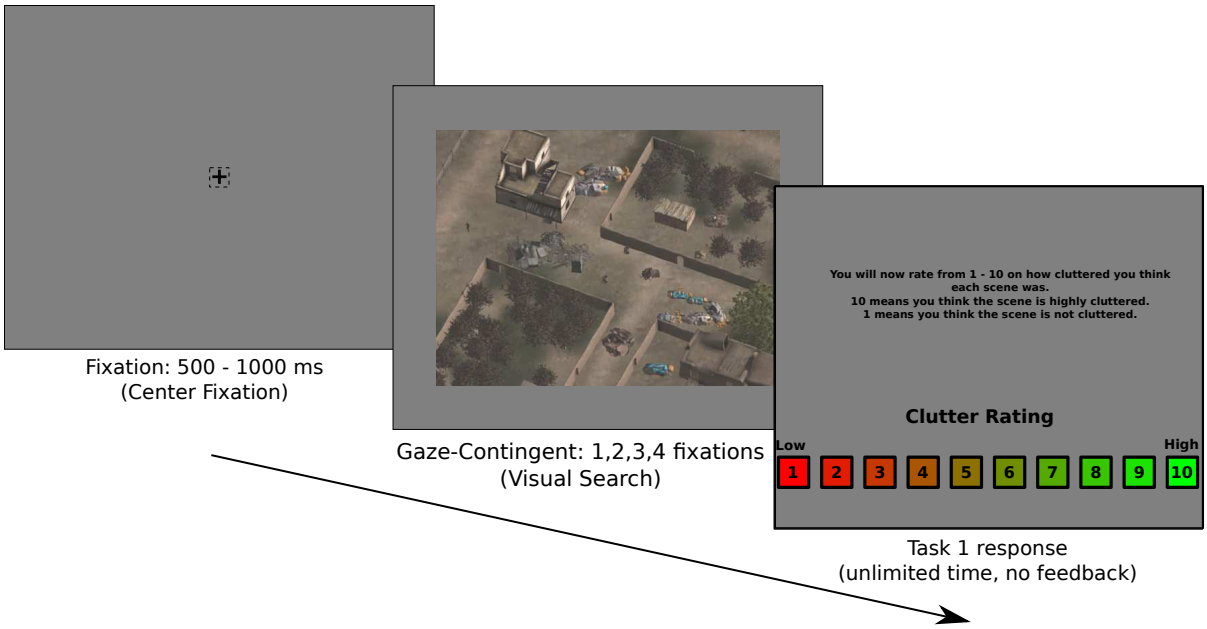


Figure 1.22: Experiment 3: Gaze-Contingent Human Clutter Judgments. Each trial begins with a fixation cross. Observers must press the spacebar while fixating at the cross and after a random interval the stimulus will appear (the same images as Experiment 2). The task however is different, and consists on rating the images on the observers perception of clutter from $\left[1-10\right]$. Observers are naive to previous visual search experiments preventing potential biases in eye movements.

1.8 Experiment 3: Gaze-Contingent Clutter Judgments

In the previous 2 experiments we showed how a foveated model is applicable to both single fixation and multi-fixation visual search. However, one caveat that we have not explored directly in our model is that we are assessing its validity with an implicit behavioural judgment such as target detectability. In this sub-section we are motivated by performing an analysis of the foveated model and the PI coefficient for clutter judgments tasks across multiple fixations, and comparing them to the regular (non-foveated) models.

As our foveated model relies on a region of interest (ROI) around the target for visual search scenarios, here we face a problem given that there is no target when an observer

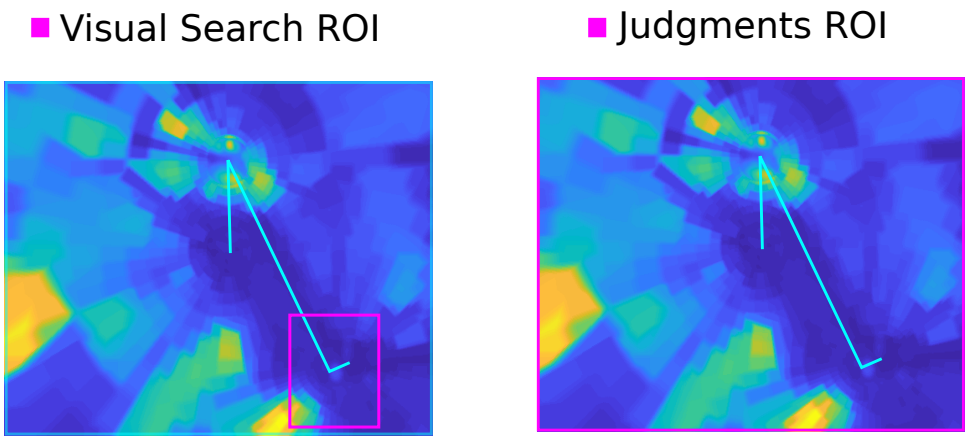


Figure 1.23: The differences in ROI (highlighted in pink) are shown contingent on the task that the observer is doing. A target ROI is placed in the location around the target of interest as the observer is doing visual search, while an image ROI captures the integration of the entire image as the observer performs a human judgment. Lines in cyan show sample saccadic trajectories.

is performing visual search. This preliminary setback seems to be a disadvantage for the foveated model as the ROI is potentially undefined. To circumvent this problem, we redefined the region of interest as the entire image under the assumption that the target *is* the stimuli. This artifice (Figure 1.23) yields strong perceptual correlations on par with the regular clutter model, and potentially suggests through our computational framework that even in 'free-viewing' conditions – observers still engage in search by nature [7].

1.8.1 Methods

A total of 6 human observers performed a human clutter judgments task, with the same collection of images shown in Experiment 2, with the exception that the observers only performed 1 session instead of 4, totalling 640 trials. The image stimuli was displayed at 1024×760 pixels resolution at 0.022 deg /px, equivalent to 22.5×16.7 d.v.a. Each trial was also gaze-contigent and terminated for any of the values of $\{1, 2, 3, 4\}$ saccades. Trials began with a center fixation for half a second until the image appeared and the

observer could freely scan the image to judge the levels of clutter in the image. At the end of each trial, the observers were asked to provide a clutter judgment from a rating of 1-10 similar to the 10 point rating of confidence of person/weapon present or absent, with the exception that the observers rated *'how cluttered they thought the scene was'*. Observers were not debriefed *a priori* with any definitions of clutter, and purely relied on their own definitions. In addition observers were also told that the stimuli would be shown for a random amount of time (as in Experiment 2). Trials in the clutter judgments task included images with the person present/absent as well as with weapons present/absent, as potentially the presence of the target may bias the judgment of clutter for the person. The stimuli used in Experiment 3 is thus identical to those of Experiment 2, but with a different task: judgments vs detection. An overview of the experimental timeline can be seen in Figure 1.22.

1.8.2 Analysis of Human Clutter Judgments

The distribution of clutter judgments across all observers can be shown in Figure 1.25 (top). Indeed, observers employ different rating strategies given that their distributions are not equal. For example, observer 'JJ' has a preference for rating most of the images within the $\left[4-6\right]$ range, while observer 'CL' has a uniform-like distribution of ratings across the full $\left[1-10\right]$ range. Recall that all observers saw the same images in a randomized order.

We later wanted to verify if the distribution of such judgments changed as a function of eye movements or number of fixations given our gaze-contingent setup. To verify that the distribution of clutter judgments did not change as eye movements increased, we ran a 2 sample Kolmogorov-Smirnov test between the human clutter scores of fixation n, and fixation $n+1$, to test if they came from the same distribution and compared them for

each observer. We found that there was no statistical difference that rejected the null hypothesis, *i.e.* all distributions are alike in shape by comparing their cumulative density function, and we found an average p value was $p > 0.87$ with $D(160) < 0.06$ within all observers. The complete per-fixation clutter judgments distribution for each observer can be seen in Figure 1.25 (bottom), where little variability in the probability density functions' shape can be visibly verified.

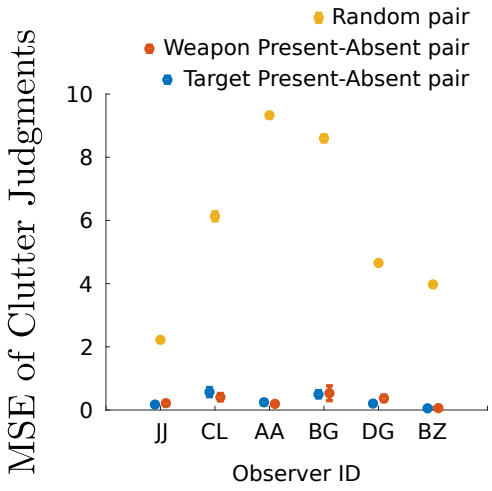


Figure 1.24: The MSE of clutter judgments.

Another aspect we evaluated with regards to the clutter judgments across observers was if they had a different perception of clutter when the target was present in the image, *vs* if the target was absent. Although the observers in this experiment were naive to Experiment 1 and 2 which required visual search, we might suspect that the target could potentially play a role in terms of an average increase in clutter rating, as the target could count as 'an additional item' [8] (or proto-object as in Yu *et al.* [20]) that contributes to set size effects in the traditional views of clutter. We thus computed the mean square error (MSE) of the clutter judgments ratings between the same image pairs: target present and target absent, and plotted these against the MSE of the ratings of two randomly selected images as a baseline. We plotted such results per observer in Figure 1.24, where we can see that not only does the person ($p < 0.0001$, t-test) not contribute to the perception of clutter in an explicit judgment task, but that the weapon ($p < 0.0001$, t-test) – which requires fine-grained discrimination, also does not contribute to the effects of explicit clutter perception.

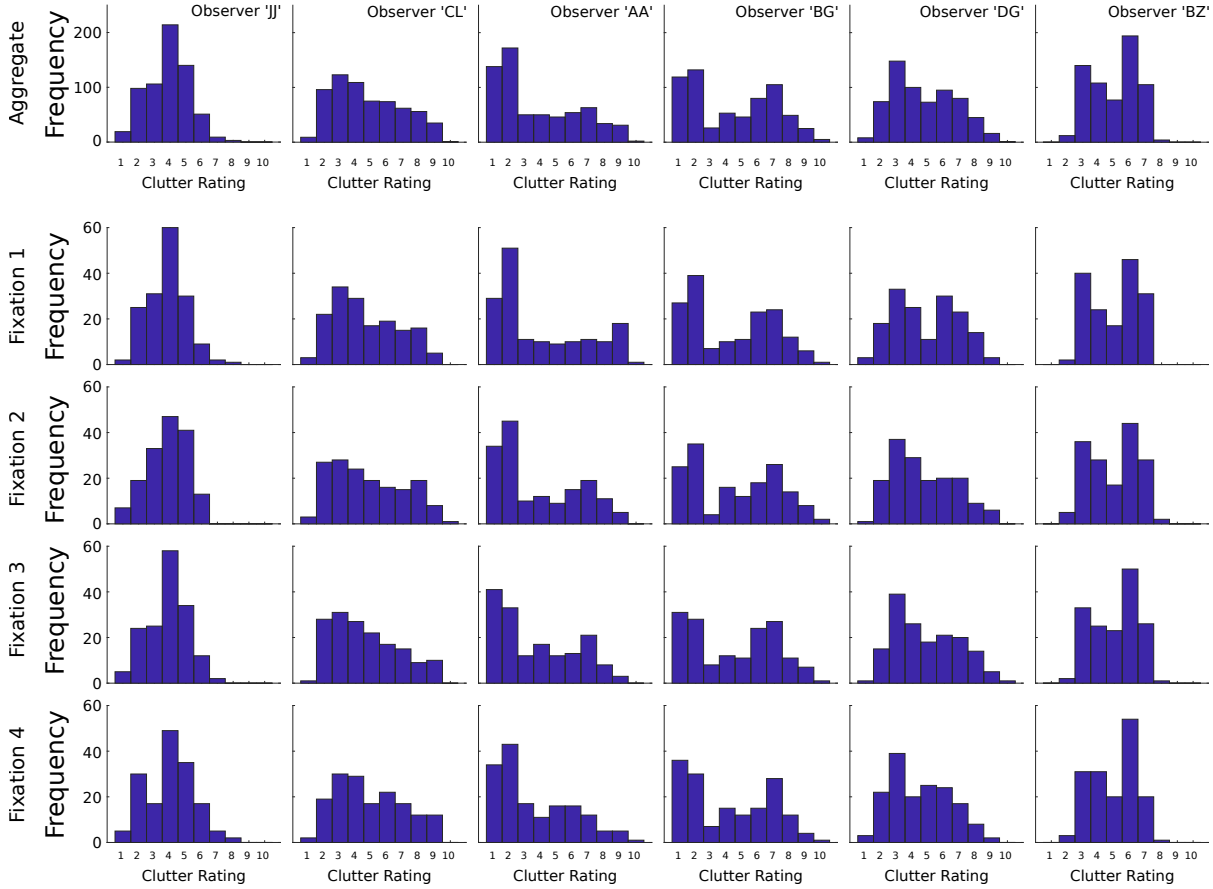


Figure 1.25: Top: A summary of the histograms of the clutter ratings across all 6 observers collapsed within fixations. Bottom: the histograms of the total amount of clutter judgments each observer made contingent on the number of eye movements (fixations). In general observers do not change their criteria as the number of eye movements increases.

1.8.3 Analysis of PI coefficient, Foveated and Regular clutter score

The first result that stems from correlating regular clutter score performance with human clutter judgments per number of eye movements is that both the Subband Entropy and ProtoObject Segmentation models correlate poorly with such ratings. Figure 1.21 shows such results. For ProtoObject Segmentation, this might be due to the nature of the model that will try to group affine superpixels together, and given that the sub selection

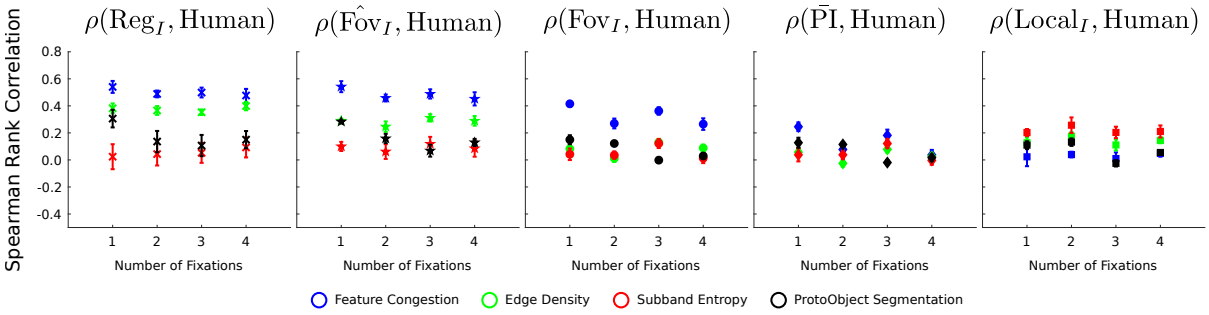


Figure 1.26: Plots of the Spearman Rank Correlation (ρ) between the different representations (Reg, $\hat{\mathrm{Fov}}$,Fov, $\bar{\mathrm{PI}}$, Local) of the clutter models and the human ratings (scale 1-10) averaged across observers and contingent on number of fixations. Here the adjusted foveated representation performs ($\hat{\mathrm{Fov}}$) on par with the regular (non-foveated) version, while the local representation (also fixation based) performs at chance. Error bars denote the standard error computed across observers.

of images in this dataset for Experiments 2 and 3 have less variability, then perhaps the grouping procedure presents a small score variance. The same is the case for Subband Entropy which depends directly on the entropy of a steerable pyramid decomposition of the image – one could say that the variance in terms of local orientation covariance is quite limited across the images. They all have terrain, walls that have similar structure that are somewhat similar under rotational invariance, as well as a potential single target. A stark contrast of Feature Congestion that is directly taking into account local covariance of color and luminance, and orientation – and Edge Density that although performing slightly below Feature Congestion, still produces significant correlations.

A closer look at Figure 1.21 from Experiment 2 (Subplots; top) verifies the advantages and limitations of these models contingent on the image statistics of our dataset. Feature Congestion and Edge Density present wide spread of scores for the regular models, while most images for Subband Entropy and ProtoObject Segmentation seem grouped into 4 and 3 clusters respectively. Perhaps the limited range of the output of these two models and the use of a limited 10 point rating scale reduces the correlations between judgments and representation of each model. Both Mack & Oliva [10], and Yu *et al.* [20], avoided

such low variability by creating a clutter rankings task and computing the median of such rankings, which notably amplifies the range for a set of ($\sim$ 100) images, implying roughly 100 unique points to perform a correlation with model outputs that decreases the likelihood of ties per image data point.

A second result is that the foveated clutter model does not perform as well as the regular clutter score. In fact, it performs somewhat between the PI coefficient which also scores low, and the regular score. However, adjusting the foveated clutter score by re-weighting the normalized PI via an area ratio of region of interest to image, shows results on par to the foveated model. The proper adjustment of the foveated clutter score will be discussed later in this subsection. The success of the adjusted foveated clutter score presents a stark contrast to the Control model (Local), where the Spearman rank correlation stays close to zero independent of number of fixations. Here, the Local model is computed by taking the mean value of the 6×6 deg ROI around the point of fixation given the dense map R. This suggests that the perception of clutter even when there is no search task, is mainly driven by global image properties such as the regular score (Reg), rather than the local region of where an observer is currently fixating at.

1.8.4 Adjustment of Foveated Clutter Score

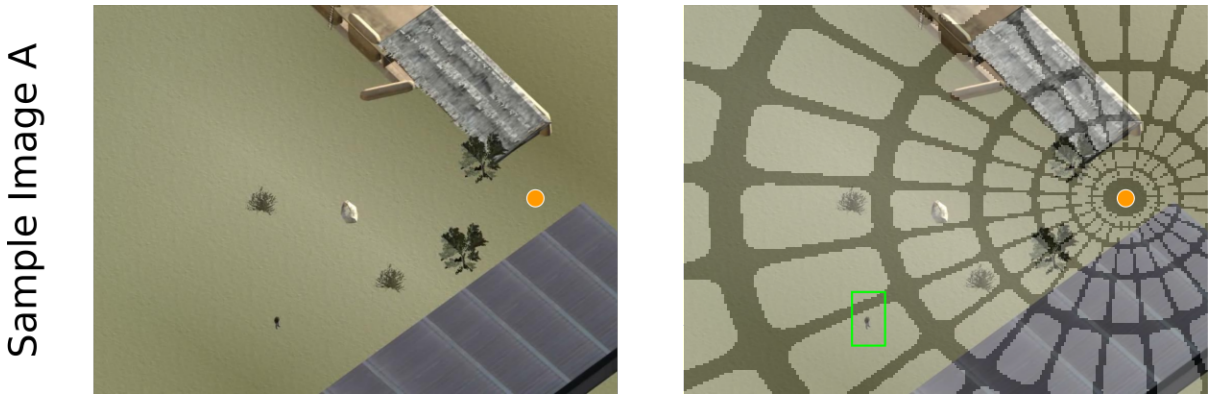
A preliminary evaluation of the foveated model yielded a weak correlations as highlighted in Figure 1.26 (middle) for the models that performed above chance such as Feature Congestion and Edge Density. We thought that one of the causes of this reduction in correlation is that the normalized PI coefficient, is compatible in range to the non-foveated regular score, such that adding the quantities near-optimally transforms the data into a new space where monotonicity and linearity of the data is enforced thus increasing the Spearman and Pearson correlations respectively. We found that when

re-computing the normalized PI, the ranges are similar, yet the effects are weakened. One possibility is that the PI should be adjusted to a smaller range, and have less of an effect as as global structure is predominant vs local information in a judgments task (Figure 1.26). However, we can not simply just 'throw away' the PI coefficient, as the general formula should still be generalizable across many conditions. One solution is to perform an adjustment of the foveated clutter score, through re-weighting the normalized PI coefficient ($\bar{\text{PI}}$) by computing the ratio of the area 6×6 deg ROI (for visual search) to the ratio of the entire image 22.5×16.7 deg (judgments)– given that the average difference is computed over more pixel elements (See Figure 1.23). This area ratio is close to ~ 0.1 – roughly an order of magnitude. This is a factor that decreases the effects of the PI, displacing it from a central figure of the model to a *nuisance* in the computation of the foveated clutter score. Attributing a meaning to such adjustment is beyond the scope of this chapter, though we suspect that the adjustment might be a surrogate for a global attention mechanism as the PI is in reality undefined given that there is no target – and the observer must covertly attend everywhere vs a small region of interest (as the target itself in the search task is also small). Thus, we have that the adjusted foveated clutter score is computed via:

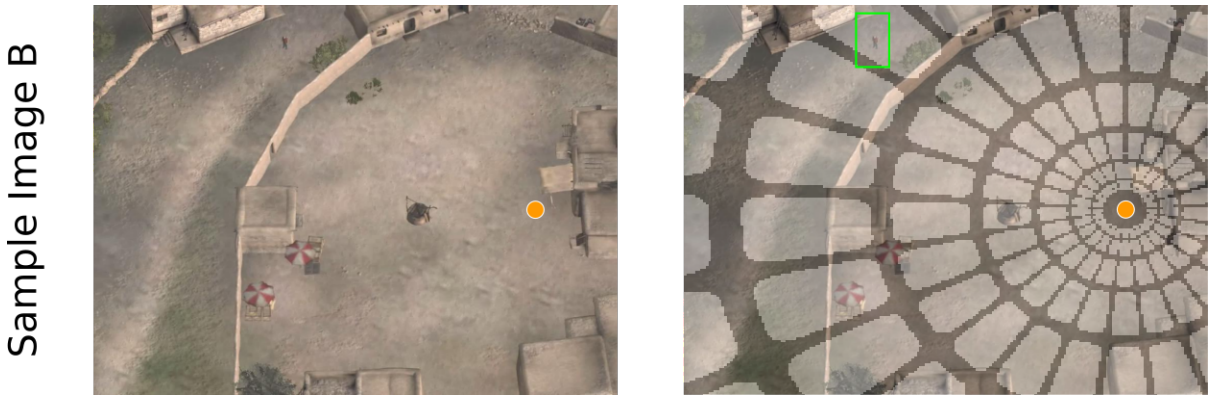
$$\hat{\text{Fov}} = \text{Reg} + \frac{\text{Area}_{ROI}}{\text{Area}_I}\bar{\text{PI}} \tag{1.32}$$

We finally tested whether the adjusted foveated clutter score, performs *on par* to the Regular model. We performed a Wilcoxon signed-rank test between the Spearman rank correlations of the 6 observers and their explicit human clutter judgments and the model scores. We found no difference for the Subband Entropy and ProtoObject Segmentation models, across all fixations, but found a significant difference for Edge Density ($p < 0.05$) for all fixations and a difference for the 2nd and 4th fixation ($p = 0.03$) for the

Feature Congestion model. But p values for the 1st and 3rd fixation yield ($p = 0.84$) and ($p = 0.22$) respectively for this last model. Thus, the irregularity we found at the *per* fixation analysis for Feature Congestion suggests that the differences we found for Feature Congestion and potentially Edge Density might be driven by a limited number of observers in our experiments, rather than model and representation differences as seen in Figure 1.26, where both models stay constant across fixations and perform well above chance (median $p = 0.03$).



(a) Sample Image A.



(b) Sample Image B.

Figure 1.27: Two image samples: A (top) and B (bottom), where a point of fixation (orange dot) is located at roughly the same retinal eccentricity away form the target. Detecting the target in image A is easier than in image B, given local effects of clutter around the target that are *pooled* within a receptive field of the peripheral architecture. The target is highlighted in green.

1.9 General Discussion

One of the main limitations of Regular clutter models is that they do not take into account the foveated nature of the human visual system, potentially incorporating effects of crowding that interfere with target detectability – specially if such behavioural metric is used as an implicit evaluation of clutter. Figure 1.27 illustrates this problem more in detail, where we have two images A and B and the target is at the same eccentricity away from the point of fixation. However, detecting the target in image A is easier than in

image B as the target is not crowded by structures of relevant features in its near vicinity, thus making it easily detectable almost from any point of fixation. We started this chapter along the previous line of thought and later developed a motivation for a foveated clutter model. The foveated clutter models takes into account both the global properties of the image and the Peripheral Integration (PI) coefficient that computes the loss of feature information due to crowding [16, 15] over a region of interest (ROI) in the visual field. In this chapter, we have directly evaluated such limitations with psychophysical data as well as designed a new foveated model that overcomes such problem.

One can thus see the limitations of non-foveated clutter models if one would like to establish that they are inversely correlated to a human behavioural metric such as target detectability. While such correlations are likely to hold as seen in the original studies of Rosenholtz *et al.* [12], and Asher*et al.*[24] – we are limited by a single data point from the set of fixations, which may lead to mis-estimating the accuracy of the clutter model. Consider the following example: a group of UX engineers are designing an interface and would like to know how easy it is for users to find a target, thus trying to determine implicitly if the image stimuli is cluttered or not. They pilot a design where observers try to find a specific target embedded in a map – a cross, analogous to the *'you are here'* sign that is commonly placed in metro maps. If the viewing time of the experiment is too long, the performance will asymptote and the validity of the model will be underestimated, yet if the viewing time is too short, the model might overestimate difficulty of search. We have seen such changes in behavior in Experiment 1 for a single fixation. Indeed, though we are using a more specific measure to quantize visual search which is number of eye movements rather than viewing time, the limitation is still evident: the designer would like to know how performance changes as a function of number of eye movements as well as their likely location given a restricted time window. A foveated model, enabled through peripheral representations as described in this chapter could enable the designer

to know realistically how likely are users to find the target. Moreover, UX-engineers could gather real eye movement data, as well as simulate fixational patterns given saliency maps or pre-specified regions of interest, to assess the likelihood of target detectability via a regression model.

Another future direction to consider when empirically testing foveated models of clutter perception is using a wider variety of stimuli with matching testing paradigms for reliable benchmarks. Most studies have relied on 1 specific dataset. For example the work of Rosenholtz *et al.* [12], used a collection of maps, the Crowding Model of van der Berg *et al.* [22], used objects from the study of Bravo & Farid [23] as well as the maps from Rosenholtz *et al.* [12]; Asher *et al.* [24], used a specific collection of scene-like imagery with very small targets analogous to ours, yet their images of scenes have a horizontal vs aerial vantage point. Figure 1.28. shows how these studies have used small collections of images of different statistical properties when analyzing clutter. The work of Mack & Oliva used 100 images and had users sub-divide them by hierarchical complexity. The stimuli of Rosenholtz *et al.* [12], totaled 25 maps with direct ranking data made by 20 human observers. The average ranking was computed as a score per image. Yu *et al.* [19], used 90 images and had 15 observers perform a ranking task where they computed the median rank as the score for each image to perform correlations analogous to Rosenholtz *et al.* [12]. Asher *et al.* [24] used 120 images (all natural scenes) for their target present/absent visual search experiment to explore behavioural correlates with clutter by optimizing over a variable width ROI. They ran their visual search experiment on 25 participants, and also had a setup where the targets were small in size (2.7 deg) – which is close to the foveal resolution limit. In our first experiment we performed our study across 13 observers and for 360 aerial images in a target detection task. Experiments 2 and 3 each had a collection of 80 images where only small targets were used, and where we collected the gaze and behavioural data

Figure 1.28: A collection of sample images showing the stark differences in the types of images used in each model and study. In general, they all differ in their nature and vantage point, ranging from natural scenes, to close-up photographs and synthetic images (ours). The first group that used a specific dataset has been placed on top of the list.

(rating/judgment) across 6 observers per experiment. Indeed, despite not having above 20 observers as most experiments, we did have a large collection of trials per observer within the 1000's, vs only having single trial data from each observer.

Indeed, as most datasets are not openly public (both the stimuli and the psychophysical data associated to each stimuli), comparisons of this nature are difficult. Given this limitation, a next step is releasing our data and behavioural outputs such that other groups may compare and evaluate their models with ours. Comparisons with standarized datasets have propelled work in computer vision as is the case of ImageNet [64] for object recognition algorithms, similarly we are looking forward to taking this step forward for research in crowding and clutter perception.

An additional difference that our study has compared to other studies, is that while some directly have a visual search or judgments task where an observer must inspect, look or scan an image yet there is no analysis on the eye-movement data. Perhaps at the time of their development, many of these clutter models were not driven by decomposing clutter by point of fixation. On the other hand, we exploited our gaze data by having a forced fixation task as done in Experiment 1 where we show that foveated models correlate stronger with target detectability than non-foveated models between and within

eccentricities. In Experiments 2 and 3 we later used the actual eye-movements to estimate both the PI coefficients as well as the foveated clutter score, in contrast to the study of van der Berg *et al.* [22], where they simulated proxy eye movements and averaged them to come up with an averaged foveated clutter score.

Despite the motivation of integrating the nature of the foveated visual system in regular clutter models, there are still some limitations to our approach. One is that our model is attention free, in that it does not compute the PI via any sort of covert attention map that may bias the detectability of a target or the perception of clutter from a top-down mechanism. In addition, our model is naive to fixation duration, as we have used the simplest of all integration rules: a cumulative min($\circ$) rule across the sequence of fixations that observers engage in for a specific task. One could imagine an updated model that weights the PI coefficient by fixation duration, which may produce stronger correlations with behavior across our experiments. Finally, we have yet to include target appearance in our model. In Experiment 1 we found that the correlations did not change by adding or removing the target in our model, but this is likely to be the case for our image stimuli where target appearance is maintained somewhat constant throughout the trials. Questions worth investigating are how does the PI coefficient vary when targets pop-out, and how do they change when they are heavily camouflaged. Indeed, there are several cases where hardly detectable or camouflaged targets may be fixated and yet observers may miss them. The stimuli required to perform such experiments where the component to control for is target appearance would be very different from the ones we (or other researchers) have used, and might be more affine to texture-heavy images, or those where visual search is highly difficult as in *'Where is Waldo'?*.

In this chapter we studied the applicability of foveated clutter models through peripheral representations when observers are engaged both in visual search and clutter judgments. As the nature of visual search is fixation-based, we believe that the the-

ory and experiments presented will provide a stepping stone for more complex uses of clutter models in human visual perception, as well as interesting applications in Human-Computer Interaction interfaces and potential machine vision models that may try to emulate the effects of visual crowding.

Chapter 2

Visual Metamers as a Generative Model of Peripheral Processing

2.1 Motivation

In the previous chapter we introduced a model that simulates the loss of perceptual information due to crowding in the visual field. In a way, we used a summary statistic (the PI coefficient) to quantize this loss and re-define the perceptual notion of clutter given a point of fixation and the task that the human observer is engaged in. Thus, we have gone from the image domain, and all of its rich representation, to a single number (or summary statistic) [47] that condenses the effects of crowding.

In this chapter, we will study the phenomena of peripheral vision via a dual approach – where rather than outputing a number that we will correlate with behaviour, we will develop a transformation that re-renders the original image, thus preserving the dimensionality and nature of the data, while finding a distribution of target images that closely matches the statistics of our original image, producing what are known as *visual metamers*[1]. Indeed, the problem of visual metamerism is defined as finding a family

of perceptually indistinguishable, yet physically different images. Thus the goal of our generative model is to find these images, to consequently evaluate them under a psychophysical paradigm to verify their connection with the physiology of V2 receptive field size as discovered earlier by Freeman & Simoncelli [1]. One of their key results was that metamerism in the visual field derives from a local texture matching procedure [65] for the multiple pooling regions in the visual field [1]. We will discuss this finding more in depth with the results given by our NeuroFovea metamer model, a foveated generative model that is based on a mixture of peripheral representations and style transfer forward-pass algorithms. Our gradient-descent free model is parametrized by a foveated VGG19 encoder-decoder which allows us to encode images in high dimensional space and interpolate between the content and texture information with adaptive instance normalization anywhere in the visual field. Our contributions include: 1) A framework for computing metamers that resembles a noisy communication system via a foveated feed-forward encoder-decoder network – We observe that metamerism arises as a byproduct of noisy perturbations that partially lie in the perceptual null space; 2) A perceptual optimization scheme as a solution to the hyperparametric nature of our metamer model that requires tuning of the image-texture tradeoff coefficients everywhere in the visual field which are a consequence of internal noise; 3) An ABX psychophysical evaluation of our metamers where we also find that the rate of growth of the receptive fields in our model match V1 for reference metamers and V2 between synthesized samples. Our model also renders metamers at roughly a second, presenting a $\times 1000$ speed-up compared to the previous work, which allows for tractable data-driven metamer experiments.

[1] These pooling regions are governed by the same equations of those used in Chapter 1 for our Foveated Clutter model.

2.2 Introduction

The history of metamers originally started through color matching theory, where two light sources were used to match a test light's wavelength, until both light sources are indistinguishable from each other producing what is called a *color metamer*. This leads to the definition of visual metamerism: when two physically different stimuli produce the same perceptual response (See Figure 2.1 for an example). Motivated by [65]'s work of local texture matching in the periphery as a mechanism that explains visual crowding, [1] were the first to create such point-of-fixation driven metamers through such local texture matching models that tile the entire visual field given log-polar pooling regions that simulate the V1 and V2 receptive field sizes, as well as having global image statistics that match the metamer with the original image. The essence of their algorithm is to use gradient descent to match the local texture ([47]) and image statistics of the original image throughout the visual field given a point of fixation until convergence thus producing two images that are perceptually indistinguishable to each other.

However, metamerism research currently faces 2 main limitations: The first is that metamer rendering faces no unique solution. Consider the potentially trivial examples of having an image I and its metamer M where all pixel values are identical except for one which is set to zero (making this difference unnoticeable), or the case where the metameric response arises from an imperceptible equal perturbation across all pixels as suggested in [66, 1]. This is a concept similar to Just Noticeable Differences ([67, 68]). However, like the work of [1, 69, 70, 65], we are interested in creating point-of-fixation driven metamers, which create images that preserve information in the fovea, yet lose spatial information in the periphery such that this loss is unnoticeable contingent of a point of fixation (Figure 2.1). The second issue is that the current state of the art for a full field of view rendering of a 512px $\times$ 512px metamer takes 6 hours for a grayscale image and roughly

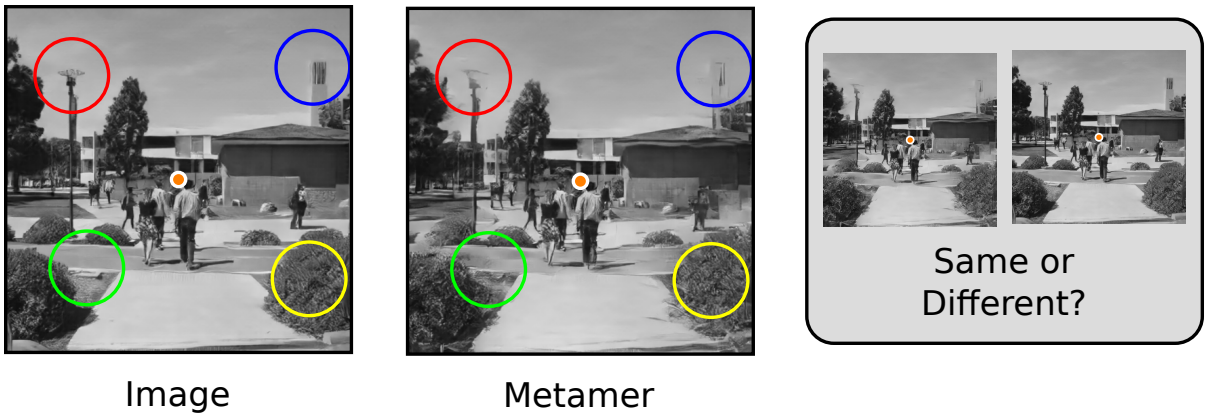


Figure 2.1: Two visual metamers are physically different images that when fixated on the orange dot (center), should remain perceptually indistinguishable to each other for an observer. Colored circles highlight different distortions in the visual field that observers do not perceive in our model.

a day for a color image. This computational constraint makes data-driven experiments intractable if they require thousands of metamers. From a practical perspective, creating metamers that are quick to compute may lead to computational efficiency in rendering of VR foveated displays and creation of novel neuroscience experiments that require metameric stimuli such as gaze-contingent displays, or metameric videos for fMRI, EEG, or Eye-Tracking.

We think there is a way to capitalize metamer understanding and rendering given the developments made in the field of *style transfer*. We know that the original model of Freeman & Simoncelli consists of a local texture matching procedure for multiple pooling regions in the visual field as well as global image content matching. If we can find a way to perform localized style transfer with proper texture statistics for all the pooling regions in the visual field, and if the metamerism via texture-matching hypothesis is correct – we can in theory successfully render a metamer.

Within the context of style transfer, we would want a complete and flexible framework where a *single* network can encode *any* style (or texture) without the need to re-train, and

with the power of producing style transfer with a single forward pass, thus enabling real-time applications. Furthermore, we would want such framework to also control for spatial and scale factors ([71]) to enable foveated pooling ([72, 46]) which is critical in metamer rendering. The very recent work of [73], provides such framework through adaptive instance normalization (AdaIN), where the content image is stylized by adjusting the mean and standard deviation of the channel activations of the encoded representation to match with the style. They achieve results that rival those of [74, 66], with the added benefit of not being limited to a single texture in a feed-forward pipeline.

In our model: we stack a peripheral architecture on top of a VGGNet ([75]) in its encoded feature space, to map an image into a perceptual space. We then add internal noise in the encoded space of our model as a characterization that perceptual systems are noisy. We find that inverting such modified image representation via a decoder results in a metamer. This breaks down our model into a foveated feed-forward 'auto' style transfer network, where the input image plays the role both of the content and the style, and internal network noise (stylized with the content statistics) serves as a proxy for intrinsic image texture. While our model uses AdaIN for style transfer and a VGGNet for texture statistics, our pipeline is extendible to other models that successfully execute style transfer and capture proper texture statistics ([76]).

2.3 Design of the NeuroFovea model

To construct our metamer we propose the following statement: A metamer M can be rendered by transferring k *localized* styles over a content image I, controlled by a set of style-to-content ratios α_i for every pooling region (i-th receptive field). More formally, our goal is to find a Metamer function $\mathbf{M}(\circ) : I \rightarrow M$, where an input image $I \in \mathbb{R}^L$ is fed through a VGG-Net encoder $\mathcal{E}(\cdot) : \mathbb{R}^L \rightarrow \mathbb{R}^D$ which is both the content and the style

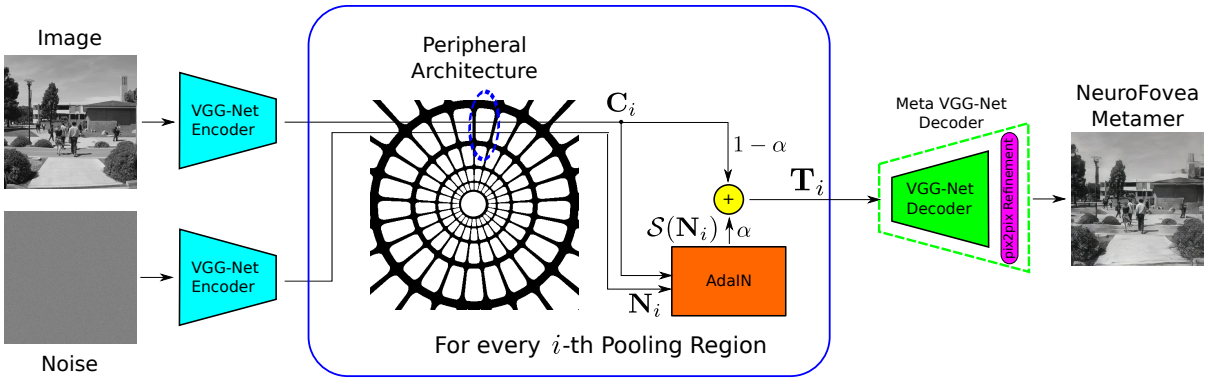


Figure 2.2: The NeuroFovea metamer generation schematic: An input image and a noise patch are fed through a VGG-Net encoder into a new feature space. Through spatial control we can produce an interpolation for each pooling region in such feature space between the stylized-noise (texture), and the content (the input image). This is how we successfully impose both global image and local texture-like constraints in every pooling region. The metamer is the output of the pooled (and interpolated) feature vector through the Meta VGG-Net Decoder.

image, to produce the content feature $\mathbf{C} \in \mathbb{R}^D$, where $\mathbf{C} = \mathcal{E}(I)$ as shown in Figure 2.2. Let $L = C \times H \times W$, and $D = C' \times H' \times W'$ where $\{C, C'\}, \{H, H'\}, \{W, W'\}$ are the image/layer channels, height, width given the convolutional structure of the encoder (we drop fully connected layers). A noise patch colored via ZCA ([77]) to match the content image's mean and variance $\mathcal{N} \sim (\mu_I, \sigma_I^2) \in \mathbb{R}^L$ is also fed through the same VGG-Net encoder producing the noise feature $\mathbf{N} \in \mathbb{R}^D$, where $\mathbf{N} = \mathcal{E}(\mathcal{N})$. This is the internal perceptual noise of the system which will later on serve us as a proxy for texture encoding. These vectors are masked through spatial control *a la* [71], and the noise is stylized via $\mathcal{S}(\cdot) : \mathbb{R}^D \to \mathbb{R}^D$ with the content which encodes the texture representation of the content in the feature space through Adaptive Instance Normalization (AdaIN). A target feature $\mathbf{T}_i \in \mathbb{R}^D$ is defined as an interpolation between the stylized noise $\mathcal{S}(\mathbf{N}_i)$ and the content $\mathbf{C}_i$ modulated by α, in the feature space $\mathbb{R}^D$ for every i-th pooling region:

$$\mathbf{T}_i(I|\mathcal{N};\alpha) = (1-\alpha)\mathbf{C}_i(I) + \alpha\mathcal{S}(\mathbf{N}_i) \tag{2.1}$$

In other words, in our quest to probe for metamerism, we are finding an intermediate representation (the convex combination) between two vectors representing the image and its texturized version (the stylized noise) in $\mathbb{R}^D$ per pooling region as seen in Figure 2.3. Within the framework of style transfer, we could think of this as a content-vs-style or structure-vs-texture tradeoff, since the style and the content image are the same. Similar interpolations have been explored in [78] via a joint pixel and network space minimization. The final target feature vector $\mathbf{T}$ is the masked sum of every $\mathbf{T}_i$ with spatial control masks w_i s.t. $\mathbf{T} = \sum w_i \mathbf{T}_i$. The metamer is the output of the Meta VGG-Net decoder $\mathcal{D}(\cdot)$ on $\mathbf{T}$, where the decoder receives only *one* vector ($\mathbf{T}$) and produces a global decoded output. Our Meta VGG-Net Decoder compensates for small artifacts by stacking a *pix2pix* [79] U-Net refinement module which was trained on the Encoder-Decoder outputs to map to the original high resolution image. Figure 2.2 fully describes our model, and the metamer transform is computed via:

$$\mathbf{M}(I|\mathcal{N};\bar{\alpha}) = \mathcal{D}(\mathcal{E}_{\sum}(I|\mathcal{N};\bar{\alpha})) = \mathcal{D}(\sum_{i=1}^{k} w_i\big[(1-\alpha_i)\mathcal{E}_i(I) + \alpha_i \mathcal{S}(\mathcal{E}_i(\mathcal{N}))\big]) \tag{2.2}$$

where $\mathcal{E}_{\sum}$ is the foveated encoder that is defined as the sum of encoder outputs over all the k pooling regions (our spatial controls masks w_i) in the visual field. Note that the decoder was not trained to generate metamers, but rather to invert the encoded image and act as $\mathcal{E}^{-1}$. It happens to be the case that perturbing the encoded representation

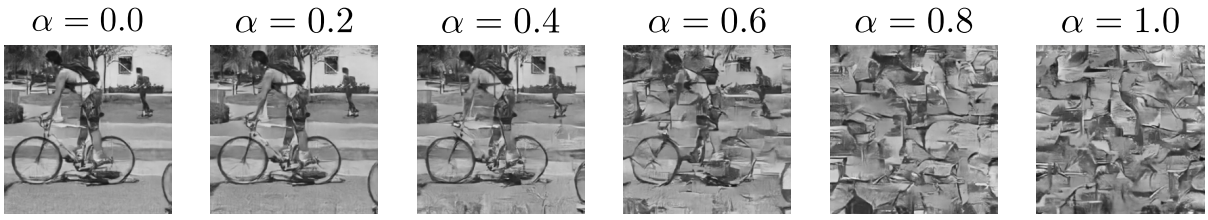


Figure 2.3: Interpolating between an image's intrinsic content and texture via a convex combination in the output of the VGG19 Encoder $\mathcal{E}$. Here we are treating the patch as a single pooling region. In our model, this interpolation given Eq. 2.1 is done for every pooling region in the visual field.

in the direction of the stylized noise by an amount specified by the size of the pooling regions, outputs a metamer. Additional specifications and training of our model can be seen in the Supplementary Material.

2.3.1 Model Interpretability

Within the framework of metamerism where distortions lie on the perceptual null space as proposed initially in color matching theory, and also in [1] for images, we can think of our model as a direct transform that is maximizing how much information to discard depending on the texture-like properties of the image and the size of the receptive fields. Consider the following: if our interpolation is projected from the encoded space to the perceptual space via P, from Eq. 2.1 we get $P\mathbf{T}_i = P(1-\alpha)\mathbf{C}_i(I) + P(\alpha)\mathcal{S}(\mathbf{N}_i)$, it follows that for each receptive field:

$$P \underbrace{\mathbf{T}_i}_{\text{metamer}} = P \underbrace{\mathbf{C}_i}_{\text{image}} + P \underbrace{\alpha(\mathcal{S}^{\perp}(\mathbf{N}_i) + \mathcal{S}^{\parallel}(\mathbf{N}_i))}_{\text{distortion}} \tag{2.3}$$

by decomposing $\mathcal{S}(\mathbf{N}_i) - \mathbf{C}_i = \mathcal{S}^{\perp}(\mathbf{N}_i) + \mathcal{S}^{\parallel}(\mathbf{N}_i)$, where $\mathcal{S}^{\parallel}$ is the projection of the difference vector on the perceptual space, and $\mathcal{S}^{\perp}(\mathbf{N}_i)$ is the orthogonal component perpendicular to such vector which lies in the perceptual null space ($P\mathcal{S}^{\perp}(\mathbf{N}_i) = \vec{0}$). The value of these components will change depending on the location of $\mathbf{C}_i$ and $\mathcal{S}(\mathbf{N}_i)$, and the geometry of the encoded space. If $||\mathcal{S}^{\parallel}(\mathbf{N}_i)||_2^2 < \epsilon$, (i.e. the image patch has strong texture-like properties), then α can vary above its critical value given that $\mathcal{S}^{\perp}(\mathbf{N}_i)$ is in the null space of P and the distortion term will still be small; but if $||\mathcal{S}^{\parallel}(\mathbf{N}_i)||_2^2 > \epsilon$, α can not exceed its critical value for the metamerism condition to hold ($P\mathbf{T}_i \approx P\mathbf{C}_i$). Thus our interest is in computing the maximal *average* amount of distortion (driven by α) given human sensitivity before observers can tell the difference. This is illustrated

in Figure 2.4 via the blue circle around $\mathbf{C}_i$ in the perceptual space which shows the *metameric boundary* for any distortion.

One can also see the resemblance of the model to a noisy communication system in the context of information theory. The information source is the image I, the transmitter and the receiver are the encoder and decoders $(\mathcal{E}, \mathcal{D})$ respectively, and the noise source is the encoded noise patch $\mathcal{E}(\mathcal{N})$ imposing texture distortions in the visual field, and the destination is the metamer M. Highlighting this equivalence is important as metamerism can also be explored within the context of image compression and rate-distortion theory as in [80]. Such approaches are beyond the scope of this paper, however they are worth exploring in future work as most metamer models purely involve texture and image analysis-synthesis matching paradigms that are gradient-descent based.

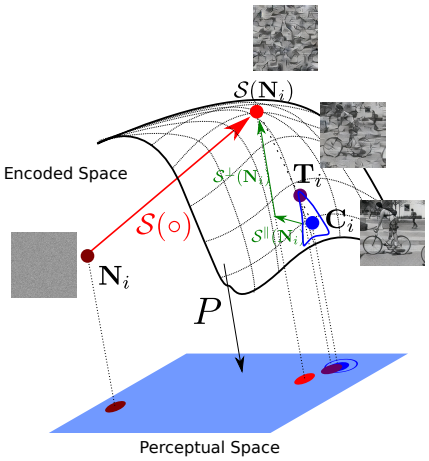


Figure 2.4: Perceptual Projection.

2.4 Hyperparameteric nature of our model

Similar to our model, the Freeman & Simoncelli model (hereto be abbreviated FS) requires a scale parameter s which controls the rate of growth of the receptive fields as a function of eccentricity. This parameter should be maximized such that an upperbound for perceptual discrimination is found. Given that texture and image matching occurs in each one of the pooling regions: a high scaling factor will likely make the image rapidly distinguishable from the original as distortions are more apparent in the periphery. Conversely, a low scaling factor might gaurantee metamerism even if the texture statistics are not fully correct given that smaller pooling regions will simulate weak effects of

crowding. Low scaling factors in that sense are potentially uninteresting – it is the value up until humans can tell the difference that is critical ([67]). FS set out to find such critical value via a psychophysical experiment where they perform the following single-variable optimization to find such upper bound:

$$s_0 = \arg\max_s \ \mathbb{E}[d'(s|\theta_{obs})] \tag{2.4}$$

s.t. $0 < d'(s|\theta_{obs}) < \epsilon$, where $d' = \Phi^{-1}(\text{HR}) - \Phi^{-1}(\text{FA})$ is the index of detectability for each observer θ_{obs}, Φ is the cumulative of the gaussian distribution, and HR and FA are the hit rate and false alarm rates as defined in [25]. However, our model is different in regards to a set of hyperparameters $\bar{\alpha}$ that we must estimate everywhere in the visual field as summarized by the γ function, where we assume α to be tangentially isotropic:

$$\alpha = \gamma(\circ; s) \tag{2.5}$$

where each α represents the maximum amount of distortion (Eq. 2.1) that is allowed for every receptive field in the visual periphery before an observer will notice. At a first glance, it is not trivial to know if α should be a function of scale, retinal eccentricity, receptive field size, image content or potentially a combination of the before-mentioned (hence the $\circ$ in the γ function's argument).

Thus, the motivation of α seems uncertain and perhaps un-necessary from the Occam's razor perspective of model simplicity. This raises the question: Why does the FS model not require any additional hyperparameters, requiring only a single scale (s) parameter? The answer lies in the nature of their model which is gradient descent based and where local texture statistics are matched for every pooling region in the visual field, while preserving global image structural information. When such condition is reached, no

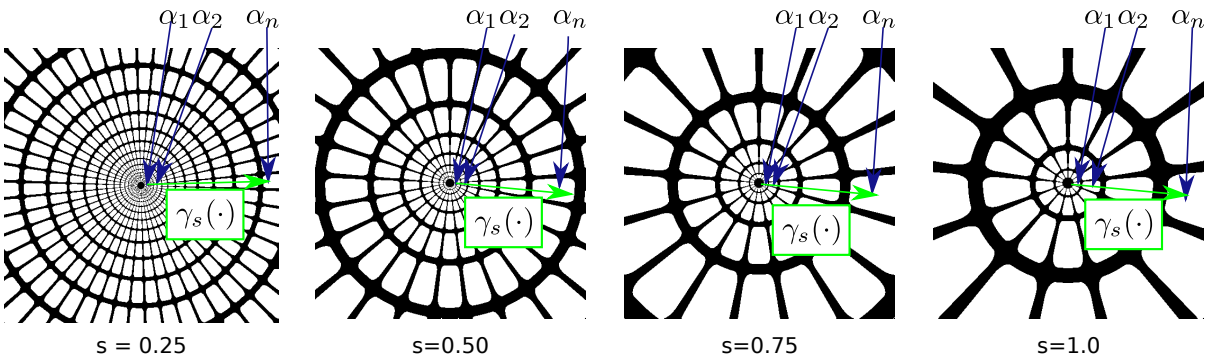


Figure 2.5: Potential issues of psychophysical intractability for the joint estimation of (s) and $\gamma(\cdot)$ as described by our model. Running a psychophysical experiment that runs an exhaustive search for upper bounds for the scale and distortion parameters for every receptive field is intractable. The goal of Experiment 1 is to solve this intractabitilty posed formally in Eq. 2.6 via a simulated experiment.

further synthesis steps are required as it is an equilibrium point. Indeed, the experiments of [81] have shown that images do not remain metameric if the structural information of a pooling region is discarded while purely retaining the texture statistics of [47]. This motivates the purpose of α where we interpolate between structural and texture representation. Thus our goal is to find that equilibirum point in one-shot, given that our model is purely feed-forward and requires no gradient-descent (Eq. 2.2). At the expense of this artifice, we run into the challenge of facing a multi-variable optimization problem that has the risk of being psychophysically intractable. Analogous to FS, we must solve:

$$s_0, \bar{\alpha_0} = \arg\max_{s,\bar{\alpha}} \ \mathbb{E}[d'(s, \bar{\alpha}|\theta_{obs})] \tag{2.6}$$

s.t. $0 < d'(s, \bar{\alpha}|\theta_{obs}) < \epsilon$. Figure 2.5 shows the potential intractability: each observer would have to run multiple rounds of an ABX experiment for a collection of many scales and α values for each location in the visual field. Consider: (S scales) $\times$ (k pooling regions) $\times$ (α_m step size for each α) $\times$ (N images) $\times$ (w trials): $SkN\alpha_m w$ trials per observer.

We will show in Experiment 1 that one solution to Eq. 2.6 is to find a relationship between each set of α's and the scale, expressed via the γ function. This requires a two stage process: 1) Showing that such γ exists; 2) Estimate γ given s. If this is achieved, we can relax the multi-variable optimization into a single variable optimization problem, where $0 < d'(s, \gamma(\circ; s)|\theta_{obs}) < \epsilon$, and:

$$s_0 = \arg\max_s \ \mathbb{E}[d'(s, \gamma(\circ; s)|\theta_{obs})] \tag{2.7}$$

2.5 Overview of Experiments

The goal of Experiment 1 is to estimate γ as a function of s via a computational simulation as a proxy for running human psychophysics. Once it is computed, we have reduced our minimization to a tractable single variable optimization problem. We will then proceed to Experiment 2 where we will perform an ABX experiment on human observers by varying the scale to render visual metamers as originally proposed by FS. We will use the images shown in Figure 2.6 for both our experiments.

Figure 2.6: A color-coded collection of images used in our experiments.

2.6 Experiment 1: Estimation of model hyperparameters via perceptual optimization

Existence and shape of γ: Given some biological priors, we would like γ to satisfy these properties:

1. $\gamma : Z \to \alpha$ s.t. $Z \in [0, \infty), \alpha \subset [0, 1)$, where $z \in Z$ is parametrized by the size (radius) of each receptive field (pooling region) which grows with eccentricity in humans.

2. γ is continuous and monotonically non-decreasing since more information should not be gained given larger crowding effects as receptive field size increases in the periphery.

3. γ has a unique zero at $\gamma(0) = 0$. Under ideal assumptions there is no loss of information in the fovea, where the size of the receptive fields asymptotes to zero.

Indeed, we found that γ is sigmoidal, and is a function of z, parametrized by s:

$$\gamma(z; s) = a + \frac{b}{c + \exp(-dz)} = -1 + \frac{2}{1 + \exp(-d(s)z)} \tag{2.8}$$

Estimation of γ: To numerically estimate the amount of α-noise distortion for each receptive field in our metamer model we need to find a way to simulate the perceptual loss made by a human observer when trying to discriminate between metamers and original images. We will define a perceptual loss $\mathcal{L}$ that has the goal of matching the distortions via SSIM of a gradient descent based method such as the FS metamers, and the NeuroFovea metamers (NF) with their reference images – a strategy similar to [82] used for perceptual rendering. We chose SSIM as it is a standard IQA metric that is monotonic with human judgements, although other metrics such as MS-SSIM and IW-SSIM show similar tuning properties for γ as shown in the Supplementary Material. Indeed the *reference* image I' for the NF metamer is limited by the autoencoder-like nature of the model where the bottleneck usually limits perfect reconstruction s.t. $I' = \mathcal{D}(\mathcal{E}(I))|_{(\alpha=0)}$, where $I' \to I$, and they are only equal if the encoder-decoder pair $(\mathcal{E}, \mathcal{D})$ allows for lossless compression. Since we can not define a direct loss function $\mathcal{L}$ between the metamers, we will need their reference images to define a convex surrogate loss function $\mathcal{L}_R$. The goal of this function should be to match the perceptual loss of both metamers *for each receptive field* k when compared to their reference images: the original image I for the FS model, and the decoded image I' for the NF model:

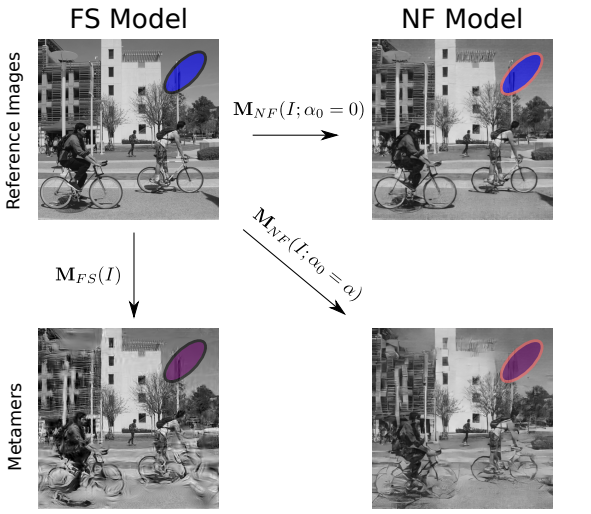


Figure 2.7: Perceptual optimization.

$$\mathcal{L}_R(\alpha|k) = \mathbb{E}(\Delta\text{-SSIM})^2 = \frac{1}{N}\sum_{j=1}^{N}(\text{SSIM}(M_{FS}^{(j,k)}, I^{(j,k)}) - \text{SSIM}(M_{NF}^{(j,k)}(\gamma_s), I'^{(j,k)}))^2 \quad (2.9)$$

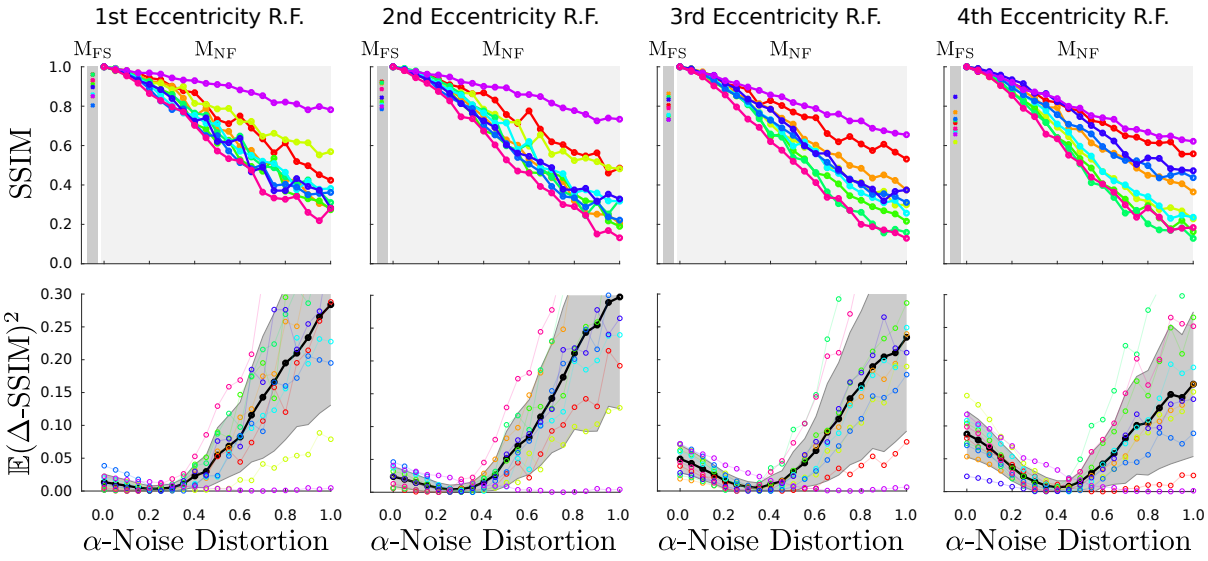


Figure 2.8: The result of each SSIM (top) for Experiment 1 for a scale of $s = 0.3$ where we find the critical α for each receptive field ring as we minimize $\mathbb{E}(\Delta\text{-SSIM})^2$ (bottom). $\mathbb{E}(\Delta\text{-SSIM})^2$ is minimized by matching the perceptual distortion of the Freeman & Simoncelli (M_{FS}) and NeuroFovea (M_{NF}) metamers in Eq. 2.9. Each color represents a different 512×512 image trajectory, the black line (bottom) shows the average. Only the first 4 eccentricity dependent receptive fields are shown.

and α_i should be minimized for each k pooling region via: $\alpha_0 = \arg\min_\alpha \mathcal{L}_R(\alpha|k)$ for the collection of N images. The intuition behind this procedure is shown in Figure 2.7. Note that if $I' = I$, *i.e.* there is perfect lossless compression and reconstruction given the choice of encoder and decoder, then the optimization is performed with reference to the same original image. This is an important observation as the reconstruction capacity of our decoder is limited despite $\mathbb{E}(\text{MS-SSIM}(I, I') = 0.86 \pm 0.04$. Only using the original image in the optimization yields poor local minima at $\alpha = 0$. Despite such limitation, we show that reference metamers can still be achieved for our lossy compression model.

Results: A collection of 10 images were used in our experiments. We then computed the SSIM score for each FS and NF image paired with their reference image across each receptive field (R.F.) and averaged those that belonged to the same retinal eccentricity. Figure 2.8 (top) shows these results, as well as the convex nature of the loss function displayed in the bottom. This procedure was repeated for all the eccentricity-dependent

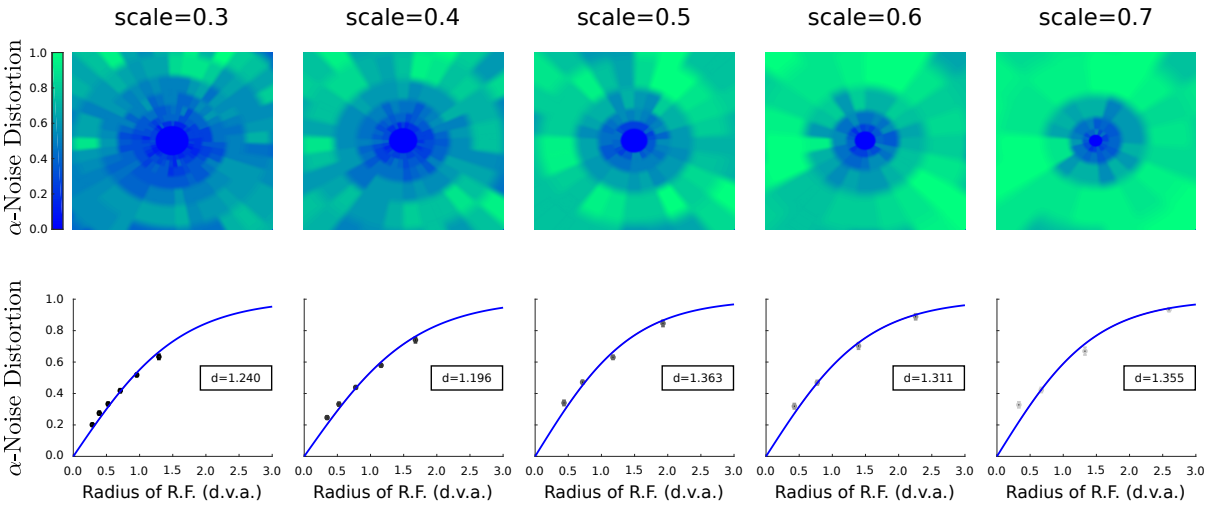
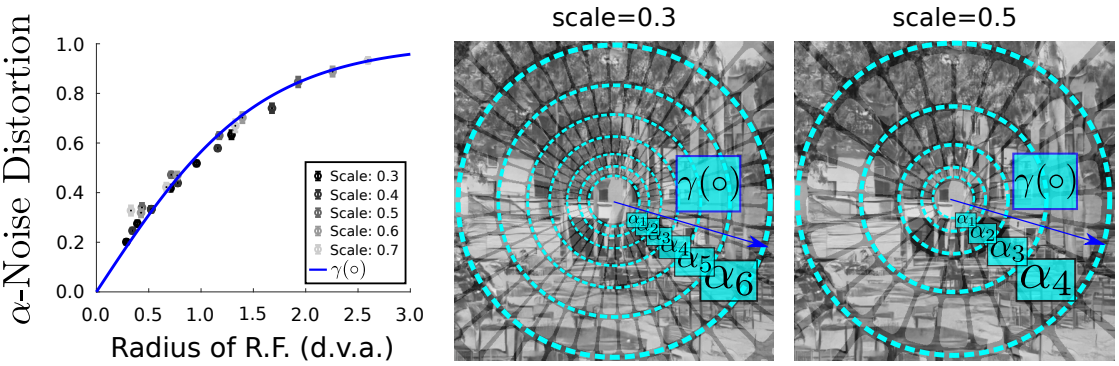


Figure 2.9: Top: The average α-noise distortion over the entire visual field for our 10 images without assuming tangential homogeneity. Notice that on average, α increases radially. Bottom: The $\gamma(\cdot)$ which completely defines the α-noise distortion for any receptive field as a function of its size (radius).

receptive fields for a collection of 5 values of scale: $\{0.3, 0.4, 0.5, 0.6, 0.7\}$. A sigmoid to estimate γ was then fitted to each α per R.F. parametrized by scale via least squares. This gave us a collection of d values that control the slope rate of the sigmoid (Eq. 2.8). These were $d : \{1.240, 1.196, 1.363, 1.311, 1.355\}$ respectively per scale, and $\{d\} = 1.281$ for the ensemble of all scales. We then conducted a 10000 sample permutation test between the pair of (z_s, α_s) points per scale and the ensemble of points across all scales $(\{z\}, \{\alpha\})$ that verified that their variation is statistically non-significant ($p \geq 0.05$). Figure 2.9 illustrates the results from such procedure. We can conclude that the parameters of γ do not vary as we vary scale. In other words, the $\alpha = \gamma(z)$ function is fixed, and the scale parameter itself which controls receptive field size will implicitly modulate the maximum α-noise distortion with a unique γ function. If the scale factor is small, the maximum noise distortion in the far periphery will be small and *vice versa* if the scale is large. We should point out that Figure 2.9 might suggest that the maximal noise distortion is contingent on image content as the scores are not uniform tangentially for the receptive

fields that lie on the same eccentricity ring. Indeed, we did simplify our model by computing an average and fitting the sigmoid. However, computing an average should approximate the maximal distortion for the receptive field size on that eccentricity in the *perceptual space* for the human observer *i.e.* the metameric boundary. We elaborate more on this idea in the discussion section.



(a) A scale invariant $\gamma(\circ)$ (b) Rendering metamers via varying s.

Figure 2.10: Metamer generation proces for Experiment 2. We modulate the distortion for each receptive field according to γ to perform an optimization as in [1].

2.7 Experiment 2: Psychophysical Evaluation of Metamerism with human observers

Given that we have estimated the value of α anywhere in the visual field via the γ function, we can now render our metamers as a function of the single scaling parameter (s), as the receptive field size z is also a function of s as shown in Figure 2.10. The psychophysical optimization procedure is now tractable on human observers and has the following form where $0 < d'(s, \gamma(z(s); s)|\theta_{obs}) < \epsilon$:

$$s_0 = \arg\max_s \ \mathbb{E}[d'(s, \gamma(z(s))|\theta_{obs})] \tag{2.10}$$

Inspired by the evaluations of [81], we wanted to test our metamers on a group of observers performing two different ABX discrimination tasks in a roving design:

1. Discriminating between Synthesized images (Synth *vs* Synth): This has been done in the original study of Freeman & Simoncelli. While this test does not gaurantee metamerism (Reference *vs* Synth), it has become a standard evaluation when probing for metamerism.

2. Discriminating between the Synthesized and Reference images (Synth *vs* Reference). This metamerism test, was not previously reported in [1] for their original images and is the most rigorous evaluation. Recently [83] argued that any model that maps an image to white noise might gaurantee metamerism under the Synth *vs* Synth condition but not against the original/reference image, thus is not a metamer.

We had a group of 3 observers agnostic to the peripheral distortions and purposes of the experiment performed an interleaved Synth vs Synth and Synth vs Reference experiment

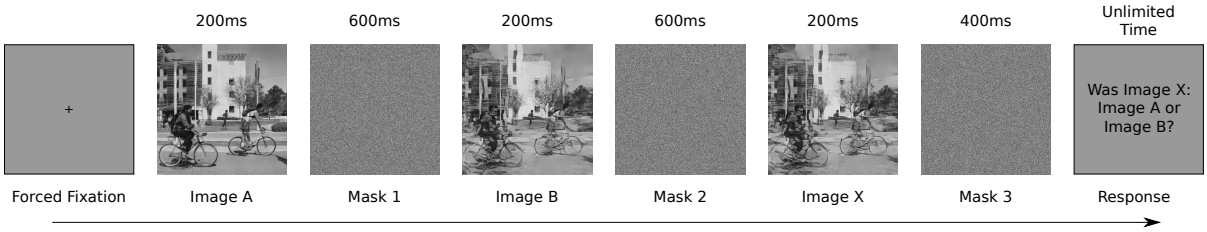


Figure 2.11: Experiment 2 shows the ABX metamer discrimination task done by the observers. Humans must fixate at the center of the image (no eye-movements) throughout the trial for it to be valid.

for NF metamers for the previous set of images (Fig. 2.6). An SR EyeLink 1000 desk mount was used to monitor their gaze for the center forced fixation ABX task as shown in Figure 2.11. In each trial, observers were shown 3 images where their task is to match the third image to the 1st or the 2nd. Each observer saw each of the 10 images 30 times per scaling factor (5) per discriminability type (2) totalling 3000 trials per observer. Images were rendered at 512×512 px, and we fixed the monitor at 52cm viewing distance and 800×600px resolution so that the stimuli subtended $26\,\text{deg} \times 26\,\text{deg}$. The monitor was linearly calibrated with a maximum luminance of $115.83 \pm 2.12\ cd/m^2$. We then estimated the critical scaling factor s_0, and absorbing factors β_0 of the roving ABX task to fit a psychometric function for Proportion Correct (PC) as in [1, 83], where the detectability is computed via $d^2(s) = \beta_0(1 - \frac{s_o^2}{s^2})\mathbb{1}_{s>s_0}$, and

$$PC(s) = \Phi\left(\frac{d^2(s)}{\sqrt{(6)}}\right)\Phi\left(\frac{d^2(s)}{2}\right) + \Phi\left(\frac{-d^2(s)}{\sqrt{(6)}}\right)\Phi\left(\frac{-d^2(s)}{2}\right) \tag{2.11}$$

Results: Absorbing gain factors β_0 and critical scales s_0 per observer are shown in Figure 2.12, where the fits were made using a least squares curve fitting model and bootstrap sampling $n = 10000$ times to produce the 68% confidence intervals. Lapse rates (λ) were also included for robustness of fit as in [84]. Analogous to Freeman & Simoncelli, we find that the critical scaling factor is 0.51 when doing the Synth vs Synth experiment which match V2, a critical region in the brain that has been identified to respond to

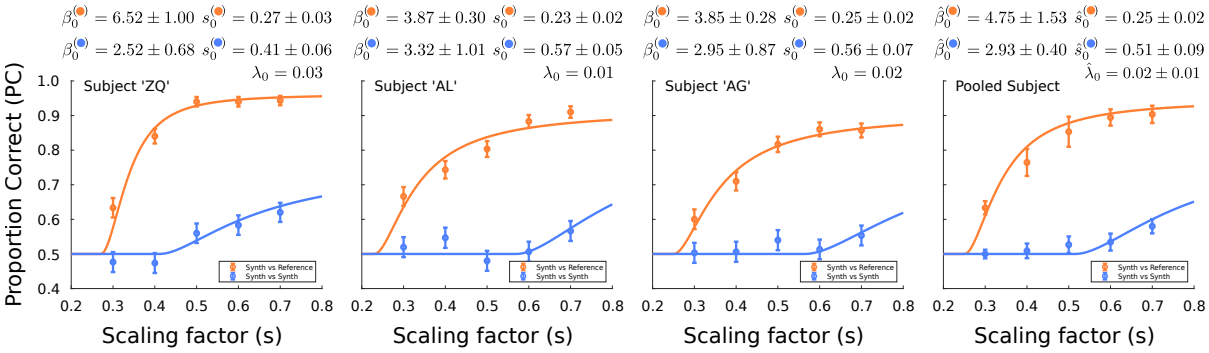


Figure 2.12: The results of the 3 observers and the pooled observer (average; shown on far right) for the Synth vs Reference and Synth vs Synth experiment for our metamers. The error bars denote the 68% confidence interval after bootstrapping the trials per observer.

texture as in [85, 86]. This suggests that the parameters we use to capture and transfer texture statistics which are different from the correlations of a steerable pyramid decomposition as proposed in [47], might the match perceptual discrimination rates of the FS metamers. This does not imply that the models are perceptually equivalent, but it aligns with the results of [76] which shows that even a basis of random filters can also capture texture statistics, thus different flavors of metamer models can be created with different statistics. In addition, we find that the critical scaling factor for the Synth vs Reference experiment is less than 0.5 (~ 0.25, matching V1) for the pooled observer as validated recently by Wallis, Funke et al., for their CNN synthesis and FS model for the Synth vs Reference condition.

2.8 Discussion

There has been a recent surge in interest with regards to developing and testing new metamer models: The SideEye model developed by [87], uses a fully convolutional network (FCN) as in [88] and learns to map an input image into a Texture Tiling Model (TTM) mongrel ([70]). Their end-to-end model is also feedforward like ours, but no use of noise is incorporated in the generation pipeline making their model fully deterministic. At first glance this seems to be an advantage rather a limitation, however it limits the biological plausilibility of metameric response as the same input image should be able to create more than one metamer. Another model which has recently been proposed is the CNN synthesis model developed by Wallis, Funke et al. (2018). The CNN synthesis model is gradient-descent based and is closest in flavor to the FS model, with the difference that their texture statistics are provided by a gramian matrix of filter activations of multiple layers of a VGGNet, rather than those used in [47].

The question of whether the scaling parameter is the only parameter to be optimized for metamerism still seems to be open. This has been questioned early in [70], and recently proposed and studied by Wallis, Funke et al. (2018), who suggest that metamers are driven by image content, rather than bouma's law (scaling factor). Figure 2.9 suggests that on average, it does seem that α must increase in proportion to retinal eccentricity, but this is conditioned by the image content of each receptive field. We believe that the hyperparametric nature of our model sheds some light into reconciling these two theories. Recall that in Figures (2.4, 2.8), we found that certain images can be pushed stronger in the direction of it's texturized version versus others given their location in the encoded space, the local geometry of the surface, and their projection in the perceptual space. This suggests that the average maximal distortion one can do is fixed contingent on the size of the receptive field, but we are allowed to *push further* (increase α) for

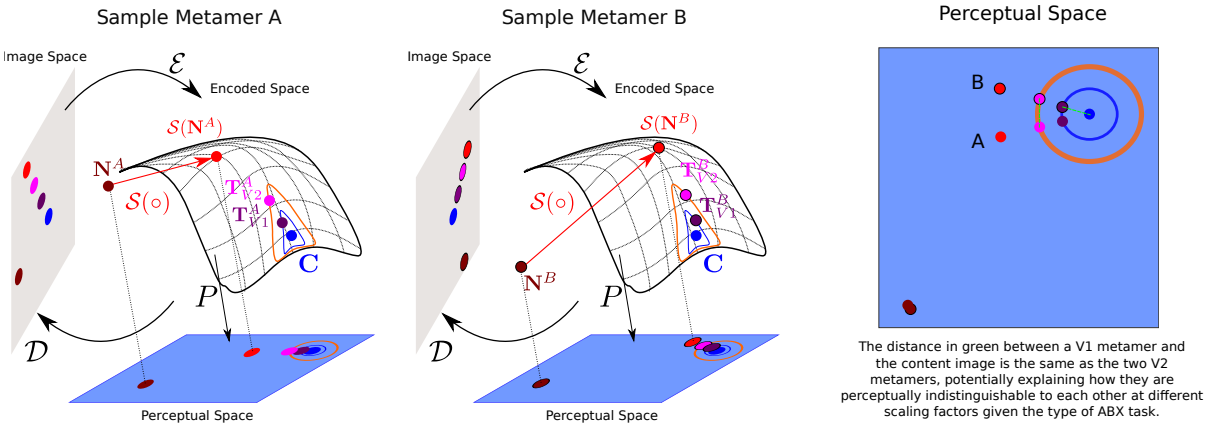


Figure 2.13: Decomposition and overview of the metamer generation process in the Image space, the Encoded space and the Perceptual space. The original image patch is coded in blue, the V1 metamers are coded in purple, and the V2 metamers are coded in pink. Dark brown represents the initial white noise that is later stylized via AdaIN through $\mathcal{S}(\circ)$. Note that these two points are far away to each other in image space, but quite closeby in perceptual space as they are also 'metameric' to each other. They are not placed on the actual encoded manifold since these points are not in the near vicinity of either $\mathbf{C}$ nor $\mathcal{S}(\mathbf{N})$, as they have no scene-like structure. The interpolation for maximal distortion is done along the line between $\mathbf{C}$ and $\mathcal{S}(\mathbf{N})$, these are the points in blue and red in the encoded space which represent the extremes of $\alpha = 0.0$ and $\alpha = 1.0$ respectively.

some images more than others, because the direction of the distortion lies closer to the perceptual null space (making this difference perceptually un-noticeable to the human observer). This is usually the case for regions of images that are periodic like skies, or grass. Figure 2.13 elaborates on how our model may potentially explain why creating synthesized samples are metameric to each other at the scales of (V1;V2), but only generated samples at the scale of V1 ($s = 0.25$) are metameric to the reference image. Here, we decompose Figure 2.4 into two separate ones for each metamer given each noise perturbation, and provide an additional visualization of the projection of the metamers in perceptual space, gaining theoretical insight on how and why metamerism arises for the synth-vs-synth condition in V2, and the synth-vs-reference condition in V1 as we demonstrated experimentally.

Our model is also different to others (FS and recently Wallis, Funke et al.) given the

role of noise in the computational pipeline. The previously mentioned models used noise as an initial seed for the texture matching pipeline via gradient-descent, while we use noise as a proxy for texture distortion that is directly associated with crowding in the visual field. One could argue that the same response is achieved via both approaches, but our approach seems to be more biologically plausible at the algorithmic level. In our model an image is fed through a non-linear hierarchical system (simulated through a deep-net), and is corrupted by noise that matches the texture properties of the input image (via AdaIN). This perceptual representation is perturbed along the direction of the texture-matched patch for each receptive field, and inverting such perturbed representation results in a metamer. Figure 2.14 illustrates such perturbations which produce metamers when projected to a 2D subspace via the locally linear embedding (LLE) algorithm ([89]). Indeed, the 10 encoded images do not fully overlap to each other and they are quite distant as seen in the 2D projection.

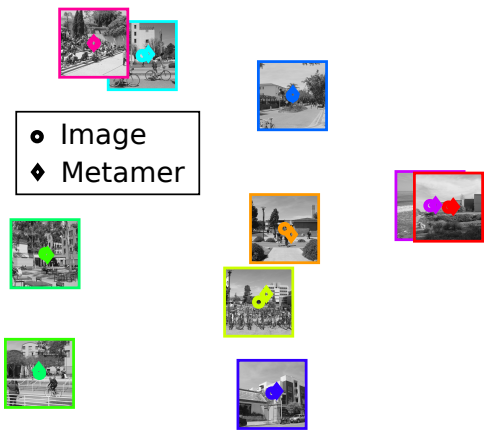


Figure 2.14: Image embeddings.

However, foveated representations when perturbed with texture-like noise seem to finely tile the perceptual space, and might act as a type of *biological regularizer* for human observers who are consistently making eye-movements when processing visual information. This suggests that robust representations might be achieved in the human visual system given its foveated nature as non-uniform high-resolution imagery does not map to the same point in perceptual space. If this holds, perceptually invariant data-augmentation schemes driven by metamerism may be a useful enhancement for artificial systems that react oddly to adversarial perturbations that exploit coarse perceptual mappings ([40, 90, 91]).

Understanding the underlying representations of metamerism in the human visual system still remains a challenge. In this chapter we propose a model that emulates metameric responses via a foveated feed-forward style transfer network. We find that correctly calibrating such perturbations (a consequence of internal noise that match texture representation) in the perceptual space and inverting such encoded representation results in a metamer. Though our model is hyper-parametric in nature we propose a way to reduce the parametrization via a perceptual optimization scheme. Via a psychophysical experiment we empirically find that the critical scaling factor also matches the rate of growth of the receptive fields in V2 ($s = 0.5$) as in Freeman & Simoncelli when performing visual discrimination between synthesized metamers, and match V1 (0.25) for reference metamers similar to Wallis, Funke et al. Finally, while our choice of texture statistics and transfer is $relu4\_1$ of a VGG19 and AdaIN respectively, our $\times$1000-fold accelerated feed-forward metamer generation pipeline should be extendible to other models that correctly compute texture/style statistics and transfer. This opens the door to rapidly generating multiple flavors of visual metamers with applications in neuroscience and computer vision.

2.9 Supplementary Material

Figure 2.15: Reference Metamers at the scale of $s = 0.25$, at which they are indiscriminable to the human observer. The color coding scheme matches the data points of the optimization in Experiment 1 and the psychophysics of Experiment 2. All images used in the experiments were generated originally at 512×512 px subtending 26×26 d.v.a (degrees of visual angle).

2.9.1 Hyperparameter search algorithm

Algorithm 2 fully describes the outline of Experiment 1.

Algorithm 2 Pipeline for Metamer hyperparameter $\gamma(\circ)$ search

procedure ESTIMATE HYPERPARAMETER: $\gamma(\circ)$ FUNCTION
Choose image dataset S_I.
Pick hyperparameter search step size α_{step}. Pick scale search step size s_{step}.
    for each image $I \in S_I$ **do**
        for each scale $s \in [s_{\text{init}} : s_{\text{step}} : s_{\text{final}}]$ **do**
            Compute baseline metamer $\mathrm{M}_{\mathrm{FS}}(I)$
            for each $\alpha \in [0 : \alpha_{\text{step}} : 1]$ **do**
                Compute metamer $\mathrm{M}_{\mathrm{NF}}(I)$
            end for
            Find the α for each receptive field that minimizes: $\mathbb{E}(\Delta\text{-SSIM})^2$.
            Fit the $\gamma_s(\circ)$ function to collection of α values.
        end for
    end for
Perform Permutation test on γ_s for all s.
    if γ_s is independent of s **then**
        $\gamma_s = \gamma$
    else
        Perform regression of parameters of γ_s as a function f of s.
        $\gamma_s = \gamma_{f(s)}$
    end if
end procedure

2.9.2 Model specificiations and training

We use $k = k_p + k_f$ spatial control windows, k_p pooling regions (θ_r receptive fields $\times$ θ_t eccentricity rings) and $k_f = 1$ fovea (at an approximate 3 deg radius). Computing the metamers for the scales of $\{0.3, 0.4, 0.5, 0.6, 0.7\}$ required $\{300, 186, 125, 102, 90\}$ pooling regions excluding the fovea where we applied local style transfer. Details regarding the decoder network architecture and training can be seen in [73]. We used the publicly available code by Huang and Belongie for our decoder which was trained on ImageNet and a collection of publicly available paintings to learn how to invert texture as well. In their training pipeline, the encoder is fixed and the decoder is trained to learn how to invert the structure of the content image, and the texture of the style image, thus when the content and style image are the same, then the decoder approximates the inverse of the encoder ($\mathcal{D} \sim \mathcal{E}^{-1}$). We also re-trained another decoder on a set of 100 images all being scenes (as a control to check for potential differences), and achieved similar outputs (visual inspection) to the publicly available one of Huang & Belongie. The dimensionality of the input of the encoder is $1 \times 512 \times 512$, and the dimensionality of the output ($relu4\_1$) is $512 \times 64 \times 64$, it is at the 64×64 resolution that we are applying foveated pooling from the initial guidance channels of the 512×512 input.

Constructions of biologically-tuned peripheral representations are explained in detail in [1, 72, 46], and are governed by the following equations:

$$f(x) = \begin{cases} cos^2(\frac{\pi}{2}(\frac{x-(t_0-1)/2}{t_0})); & -(1+t_0)/2 < x \leq (t_0-1)/2 \\ 1; & (t_0-1)/2 < x \leq (1-t_0)/2 \\ -cos^2(\frac{\pi}{2}(\frac{x-(1+t_0)/2}{t_0})) + 1; & (1-t_0)/2 < x \leq (1+t_0)/2 \end{cases} \tag{2.12}$$

$$h_n(\theta) = f\Big(\frac{\theta - (w_\theta n + \frac{w_\theta(1-t_0)}{2})}{w_\theta}\Big); w_\theta = \frac{2\pi}{N_\theta}; n = 0, ..., N_\theta - 1 \tag{2.13}$$

$$g_n(e) = f\Big(\frac{\log(e) - [\log(e_0) + w_e(n+1)]}{w_e}\Big); w_e = \frac{\log(e_r) - \log(e_0)}{N_e}; n = 0, ..., N_e - 1 \tag{2.14}$$

where $f(x)$ is a cosine profiling function that smoothes a regular step function, and $h_n(\theta)$,$g_n(e)$, are the averaging values of the pooling region w_i at a specific angle θ and radial eccentricity e in the visual field. In addition we used the default values of visual radius of $e_r = 26\,\text{deg}$, and $e_0 = 0.25\,\text{deg}$ [2], and $t_0 = 1/2$. The scale s defines the number of eccentricities N_e, as well as the number of polar pooling regions N_θ from $\langle 0, 2\pi]$. We perform the foveated pooling operation on the output of the Encoder. Since the encoder is fully convolutional with no fully connected layers, guidance channels can be used to do localized (foveated) style transfer.

Our pix2pix U-Net refinement module took 3 days to train on a Titan X GPU, and was trained with 64 crops (256×256) per image on 100 images, including horizontally mirrored versions. We ran 200 training epochs of these 12800 images on the U-Net architecture proposed by [79] which preserves local image structure given an adversarial and L2 loss.

[2] We remove central regions with an area smaller than 100 pixels, and group them into the fovea

2.9.3 Metamer Model Comparison

The following table summarizes the main similarities and differences across all current models:

Model	FS (2011)	CNN-Synthesis (2018)	SideEye (2017)	NF (Ours)
Feed-Forward	-	-	✓	✓
Input	Noise	Noise	Image	Image
Multi-Resolution	✓	✓	-	-
Texture Statistics	Steerable Pyramid	VGG19 *conv*-$1_1, 2_1, 3_1, 4_1, 5_1$	Steerable Pyramid	VGG19 *relu*4_1
Style Transfer	Portilla & Simoncelli	Gatys et al.	Rosenholtz et al.	Huang & Belongie
Foveated Pooling	✓	✓	(Implicit via FCN)	✓
Decoder (trained on)	-	-	metamers/mongrels	images
Moveable Fovea	✓	✓	✓	✓
Use of Noise	Initialization	Initialization	-	Perturbation
Non-Deterministic	✓	✓	-	✓
Direct Computable Inverse	-	-	(Implicit via FCN)	✓
Rendering Time	hours	minutes	miliseconds	seconds
Image type	scenes	scenes/texture	scenes	scenes
Critical Scaling (*vs* Synth)	0.46	$\sim \{0.39/0.41\}$	Not Required	0.5
Critical Scaling (*vs* Reference)	Not Available	$\sim \{0.2/0.35\}$	Not Required	0.24
Experimental design	ABX	Oddball	-	ABX
Reference Image in Exp.	Metamer	Original	-	Compressed via Decoder
Number of Images tested	4	400	-	10
Trials per observers	~ 1000	~ 1000	-	~ 3000

Table 2.1: Metamer Model comparison

Encoding: $\mathcal{E}(M) = \mathcal{E}(I)$
Decoding: $M = \underset{N}{\operatorname{argmin}} \, ||\mathcal{E}(I) - \mathcal{E}(N)||$

Encoding: $\mathcal{E}(M) = \mathcal{E}(I) + N$
Decoding: $M = \mathcal{D}(\mathcal{E}(I) + N)$

Original | Freeman & Simoncelli | Reference | NeuroFovea

Figure 2.16: Algorithmic (top) and visual (bottom) comparisons between our metamers and a sample from Freeman & Simoncelli for a scaling factor of 0.3. Each model has it's own limitations: The FS model can not directly compute an inverse of the encoded representation to generate a metamer, requiring an iterative gradient descent procedure. Our NF model is limited by the capacity of the encoder-decoder architecture as it does not achieve lossless compression (perfect reconstruction).

2.9.4 Pilot Experiments

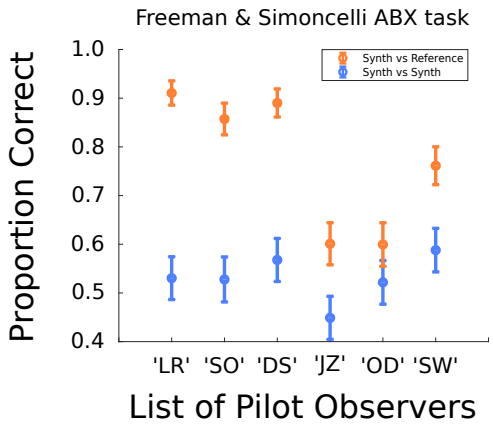


Figure 2.17: Pilot Data on FS metamers.

In a preliminary psychophysical study, we ran an experiment with a collection of 50 images and 6 observers on the FS metamers. Observers performed a single session of 200 trials of the FS metamers where the scale was fixed at $s = 0.5$. We found the following: While we found that the synthesized images were metameric to each other for the scaling factor of 0.5, the FS metamers were not metameric to their reference high-quality images at the scale of 0.5. Only a sub-group of observers: *'LR'*,*'SO'*,*'DS'* scored well above chance in terms of discriminating the images in the ABX task. These results are in synch with the evalutions done by Wallis, Funke et al., which varied scale and found a critical value to be less than 0.5 and rather closer to 0.25 within the range of V1.

2.9.5 Estimation of Lapse-rate (λ) per observer

The motivation behind estimating the lapse rate is to quantify how engaged was the observer in the experiment, as well as providing a robust estimate of the parameters in the fit of the psychometric functions. Not accounting for lapse rate may dramatically affect the estimation of these parameters as suggested in [84]. In general lapse rates are computed by penalizing a psychometric function $\psi(\circ)$ that ranges between some lower bound and upper bound usually $[0, 1]$. To estimate the lapse rate λ, a new $\psi'(\circ)$ is defined to have the following form:

$$\psi'(\circ) = b + (1 - b - \lambda)\psi(\circ) \tag{2.15}$$

Recall that for us, our psychometric fitting function $\psi(\circ) = PC_{ABX}(s)$ is defined by Equation 2.11 and parametrized by both the absorbing factor β_0 and the critical scaling factor s_0:

$$PC_{ABX}(s) = \Phi\left(\frac{d^2(s)}{\sqrt{(}6)}\right)\Phi\left(\frac{d^2(s)}{2}\right) + \Phi\left(\frac{-d^2(s)}{\sqrt{(}6)}\right)\Phi\left(\frac{-d^2(s)}{2}\right) \tag{2.16}$$

where we have:

$$d^2(s) = \beta_0(1 - \frac{s_o^2}{s^2})\mathbb{1}_{s>s_0} \tag{2.17}$$

To compute the new $\psi'(\circ)$, we notice first that our ψ is bounded between $[0.5, 1]$, and that the new ψ' will be a linear combination of a correct guess for a lapse, and a correct decision for a non-lapse from which we obtain:

$$PC(s) = \lambda + (1 - 2\lambda)PC_{ABX}(s) \tag{2.18}$$

as derived in [92] which includes lapse rates for an AXB task. When fitting the curves for each of the $n = 10000$ bootstrapped samples, we restricted the lapse rate to vary between $\lambda = [0.00, 0.06]$ as suggested in [84], and found the following lapse rates:

Observer 1: $\lambda_{ZQ}^{RS} = 0.0248 \pm 0.0209$, $\lambda_{ZQ}^{SS} = 0.0430 \pm 0.0228$.

Observer 2: $\lambda_{AL}^{RS} = 0.0008 \pm 0.0062$, $\lambda_{AL}^{SS} = 0.0166 \pm 0.0215$.

Observer 3: $\lambda_{AG}^{RS} = 0.0141 \pm 0.0243$, $\lambda_{AG}^{SS} = 0.0218 \pm 0.0236$.

We later averaged these lapse rates as there is an equal probability of each type of trial to appear (Synth vs Synth, or Reference vs Synth), and refitted each curve with the new pooled lapse rate estimates λ'. Indeed, each observer did both experiments in a roving paradigm, rather than doing one experiment after the other – thus we should only have *one* estimate for lapse rate per observer. It is worth mentioning that re-performing the fits with separate lapse rates did not significantly affect the estimates of critical scaling values, as one might argue that higher lapse rates will significantly move the critical scaling factor estimates. This is not the case as the absorbing factor β does not place an upper bound for the psychometric function at 1.

Our critical estimates of lapse rates were: $\lambda_{ZQ} = 0.0339$, $\lambda_{AL} = 0.0087$, $\lambda_{AG} = 0.0179$, as shown in Figure 2.12.

The estimates (critical scale (s_0), absorbing factor (β_0) and lapse rate (λ_0)) shown for the pooled observer were obtained by averaging the estimates over the 3 observers.

2.9.6 Robustness of estimation of γ function

In this subsection we show how the perceptual optimization pipeline is robust to a selection of IQA metrics such as MS-SSIM (multi-scale SSIM [3]) from [93] and IW-SSIM (information content weighted SSIM) from [94].

There are 3 key observations that stem from these additional results:

1. The sigmoidal natural of the γ function is found again and is also scale independent, showing the broad applicability of our perceptual optimization scheme and how it is extendable to other IQA metrics that satisfy SSIM-like properties (upper bounded, symmetric and unique maximum).

2. The tuning curves of MS-SSIM and IW-SSIM look almost identical, given that IW-SSIM is not more than a weighted version of MS-SSIM where the weighting function is the mutual information between the encoded representations of the reference and distortion image across multiple resolutions. Differences are stronger in IW-SSIM when the region over which it is evaluated is quite large (*i.e.* an entire image), however given that our pooling regions are quite small in size, the IW-SSIM score asymptotes to the MS-SSIM score. In addition both scores converge to very similar values given that we are averaging these scores over the images and over all the pooling regions that lie within the same eccentricity ring. We found that $\sim 90\%$ of the maximum α's had the same values given the 20 point sampling grid that we use in our optimization. Perhaps a different selection of IW hyperparameters (we used the default set), finer sampling schemes for the optimal value search, as well as averaging over more images, may produce visible differences between both metrics.

[3] scale in the context of SSIM is referred to resolution (as in scales of a laplacian pyramid), and is not to be confused with the scaling factor s of our experiments which encode the rate of groth of the receptive fields.

3. The sigmoidal slope is smaller for both IW-SSIM and MS-SSIM *vs* SSIM, which yields more conservative distortions (as α is smaller for each receptive field). This implies that the model can still create metamers at the estimated found scaling factors of 0.21 and 0.50, however they may have different *critical* scaling factors for the reference vs synth experiment, and for the synth vs synth experiment. Future work should focus on psychophysically finding these critical scaling factors, and if they still are within the range of rate of growth of receptive field sizes of V1 and V2.

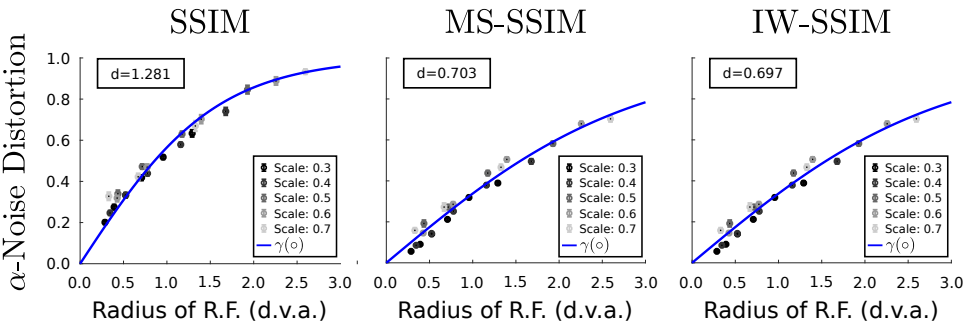
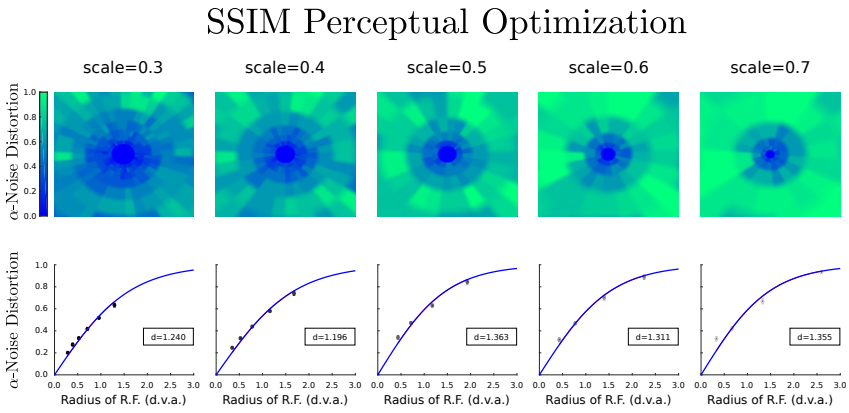
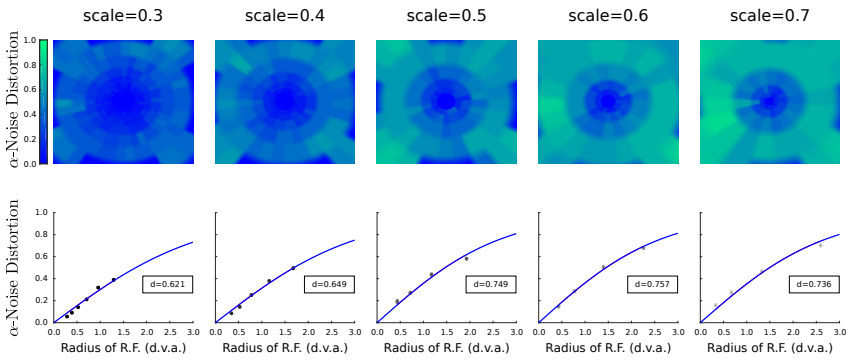


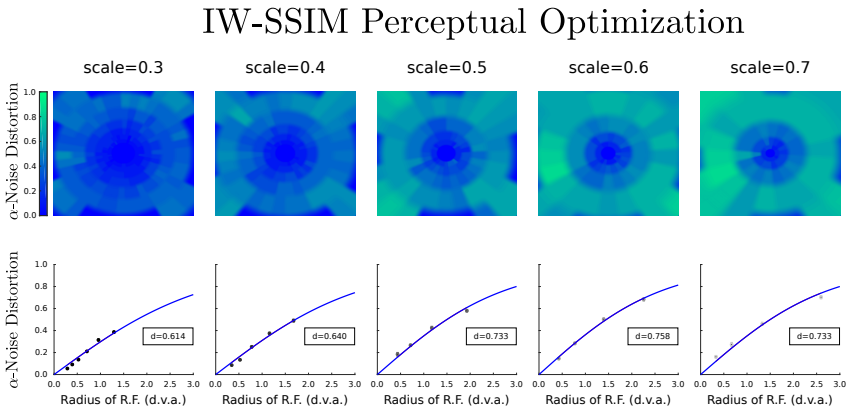
Figure 2.18: A collection of scale invariant $\gamma(\circ)$'s across multiple IQA metrics for the perceptual optimization scheme of Experiment 1. In this figure we superimpose all maximal α-noise distortions for each scale, and find a function that fits all the points showing that γ is indepedent of scale.



(a) Perceptual Optimization with SSIM.



(b) Perceptual Optimization with MS-SSIM.



(c) Perceptual Optimization with IW-SSIM.

Figure 2.19: Top: The maximum α-noise distortion computed per pooling region, and collapsed over all images for each IQA metric. Bottom: When averaging across all the pooling regions for each retinal eccentricity, we find that the γ function is invariant to scale as in our original experiment – suggesting that our perceptual optimization scheme is flexible across IQA metrics.

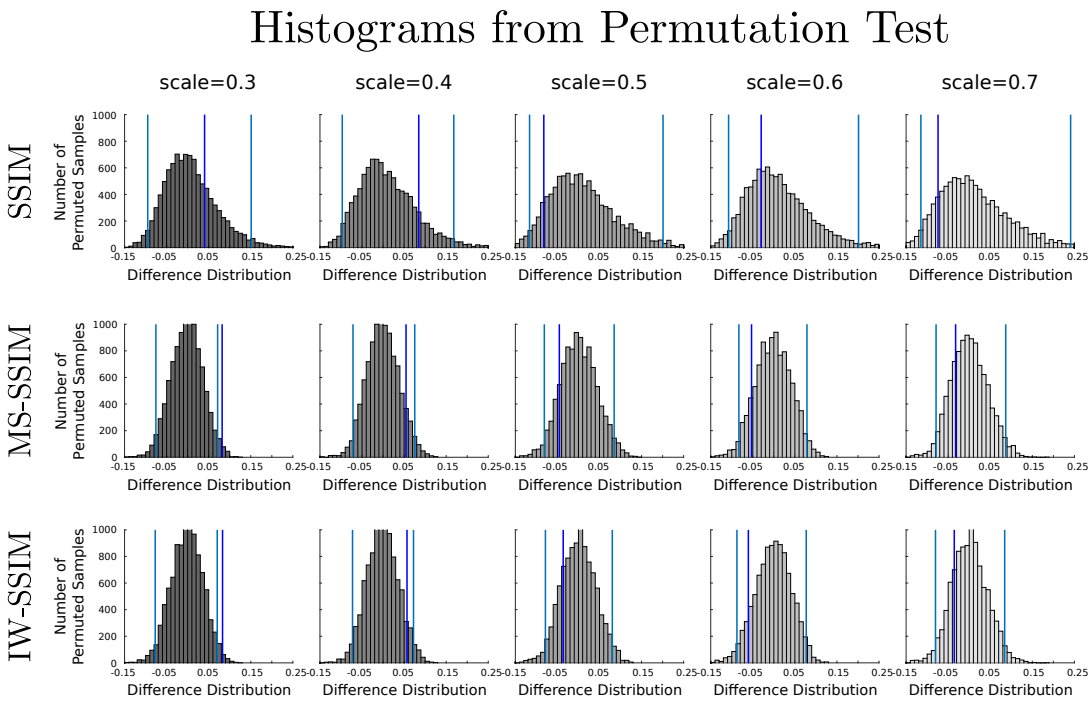


Figure 2.20: A permutation test was ran and determined that each γ function is also scale independent under the 99% confidence interval (CI), as we increased the CI to account for false discovery rates (FDR). Indeed, when we perform the permutation tests and use a 95% confidence interval (shown in the figure with the vertical lines in cyan), all curves except for MS-SSIM and IW-SSIM only for the scaling factor of 0.3 show a significant difference $p \sim 0.02$ (non FDR-corrected), potentially due to small receptive field sizes, which bias the estimates. All other differences in the d parameter of the sigmoid function, with respect to the average fitted sigmoid, are statistically insignificant.

Chapter 3

Exploiting the limitations of peripheral processing via machine assisted visual search

3.1 Motivation

In the two previous chapters we: 1) modelled the foveated nature of the human visual system and its relevance to the perception of clutter and search; 2) developed a new metamer model that matched the distortions of the human field of view – mainly the visual periphery. Given that machines (compared to humans), do not face these same limitations: in particular a foveated field of view, as well as cognitive load in decision making, the focus of this chapter is to develop and test two novel systems that assist humans when engaging in search via such computer-based (machine) assistance.

The first system is a feedback-enabled cognitive optimizer which we will call the *attention allocation aid (AAAD)*, which uses real-time physiological data to improve human performance in a realistic sequential visual search task. Specifically, using experimental

eye-tracking data and measurements about target detectability across the human visual field, we develop functional models of detection accuracy as a function of search time, number of eye movements, scan path, and image clutter. These models are then used by the AAAD in conjunction with real time eye position data to make probabilistic estimations of attained search accuracy and to recommend that the observer either move on to the next image or continue exploring the present image. In this chapter, we provide an experimental evaluation of a scenario motivated from human supervisory control in surveillance missions, where we find a $\times 1.5$ factor reduction in visual search time per trial, confirming the time-efficiency benefits of the AAAD system while preserving accuracy.

The second system is motivated by the advent of modern expert systems driven by deep learning that supplement human experts when engaging in visual search (e.g. radiologists, dermatologists, surveillance scanners). Thus, we analyze how and when do such expert systems enhance human performance in a fine-grained small target visual search task, and if such benefits are similar or complementary to those of the previous system. We set up a 2 session factorial experimental design in which humans visually search for a target with and without a Deep Learning (DL) expert system. We then evaluate human changes of target detection performance and eye-movements in the presence of the DL system. We find that performance improvements with the DL system (computed via a Faster R-CNN with a VGG16) interacts with observer's perceptual abilities (e.g., sensitivity). The main results include: 1) The DL system reduces the False Alarm rate per Image on average across observer groups of both high/low sensitivity; 2) *Only* human observers with high sensitivity perform better than the DL system, while the low sensitivity group does not surpass individual DL system performance, even when aided with the DL system itself; 3) Increases in number of trials and decrease in viewing time were mainly driven by the DL system only for the low sensitivity group. 4) The DL system aids the human observer to fixate at a target by the 3rd fixation, potentially explaining

boosts in performance. These results provide insights of the benefits and limitations of deep learning systems that are collaborative or competitive with humans.

3.2 Introduction

Visual search is an ubiquitous activity that humans engage in every day for a multitude of tasks. Some of these search scenarios are explicit such as: searching for our keys on our desk; while other are implicit such as looking for pedestrians on the street while driving [7]. Visual search may also be trivial as in the previous example or may require stronger degrees of expertise accumulated even over many years such as radiologists searching for tumors in mamograms, as well as military surveillance operators, or TSA agents who must go over a high collection of images in the shortest amount of time. This need has developed a surge of machine-assisted systems, also due in part to the high amount of data that is available across multiple applications ranging across medicine, military, homeland security and e-commerce, enabling machine learning driven expert systems to learn the visual representations that rival human observers. However, as of the present day, it is ultimately the job of a human to ensure that this data is processed both quickly and accurately, independent of having any form of machine-based assistance.

For example, *supervisory systems* involving collaboration between human operators and unmanned vehicles often require the sequential processing of imagery that is generated by the autonomous vehicles' on-board cameras for the purpose of finding targets, analyzing terrain, and making key planning decisions [95]. The incredible volume of data generated by modern sensors, combined with the complex nature of modern mission scenarios, makes operators susceptible to information overload and attention allocation inefficiencies [96], which can lead to detrimental performance and potentially dire consequences [97]. As such, the development of tools to improve human performance in visual data analysis tasks is crucial to ensuring mission success.

Similar man-machine collaborative efforts driven by deep learning and computer vi-

sion systems have been performed in the medical imaging domain, more specifically in radiology. Litjens *et al.* [98] compiled an overview of 300 Deep Learning (DL) papers applied to medical imaging. In the work of Kooi *et al.*, CNN's and other Computer Aided Detection and Diagnosis (CAD) classifiers are compared to each other as automatic diagnosis agents [99]. They find that deep learning systems rival expert radiologists, as is the recent paper of Rajpurkar *et al.* when having radiologists diagnosing pneumonia [100]. Arevalo *et al.* benchmark CNN's to classical computer vision models such as HOG and explore the learned representations by such deep networks in the first convolutional layer [101]. The majority of studies have evaluated automated intelligent agents via classical computer vision or end-to-end deep learning architectures *v.s.* humans.

Other bodies of work regarding collaborative human-machine scenarios in computer vision tasks include: image annotation [102], machine teaching [103, 104], visual conversational agents [105], cognitive optimization [106], and fined-grained categorization [107]. Conversely, there has also been a recent trend comparing humans against machines in certain tasks with the goal of finding potential biological constraints that are missing in deep networks. These comparisons have been done in object recognition [108, 5, 13], perceptual discrimination [109] and visual attention [110]. Moreover, there are many applications where mixed DL and human teams are a likely next step prior to replacement of the human expert by the expert system [111, 112, 106, 37, 95]. Given current paradigms in computer vision technology that rely on bounding box candidate regions proposals and evaluations of multiple regions of interest [113] as is the case of models from HOG [114] and DPM [115] to Faster R-CNN [9] and YOLO [39], how well do they integrate with humans whose visual search system is foveated by nature [46, 116, 117]?

Attention allocation aids have been studied in the context of human supervisory control of large data acquired by multiple automated agents (e.g., [118, 96]). Such scenarios present the challenge of a human having to inspect large data sets with possible errors

due to visual limitations, attentional bottlenecks, and fatigue. The use of advanced physiological sensing through eye-tracking technology has become a viable option for both the assessment of the operator cognitive state, and the evaluation of operator performance in a number of realistic applications, e.g. [119]. One line of research attempts to use eye-tracking measurements to detect physiological and cognitive precursors to behavior such as perceived workload, fatigue, or situational awareness. Another line of research has focused on how to augment human capabilities in coordinating multiple tasks. For example, models have been used to optimize how observers split their attentional resources when simultaneously conducting two different visuo-cognitive tasks [120]. Indeed, objective measures such as blink rates [121], pupil diameter [122, 123] , and fixation/saccade characteristics [124], all have correlations to cognitive processing, although the use of such measurements as reliable indicators of operator mental states is not fully understood [125]. If undesirable states can be accurately anticipated with physiological measures, then they can be used to drive automated aids that mediate operator resources through, e.g., optimization of task schedules [126] or adaptive automation schemes [127, 128].

The first part of this chapter focuses on the development and experimental verification of a novel, attention allocation aid that is designed to help human operators in a sequential visual search task, which requires the detection and classification of targets within a simulated landscape. Our study is primarily motivated by surveillance applications that require humans to *efficiently* detect and classify high value targets within videos generated by remote sensors, e.g., mounted on unmanned vehicles; however, the presented method is applicable to a variety of application domains, that present similar detectability rates.

In the second part of this chapter we are interested in evaluating the influences of a Deep Learning (DL) system on human behavior *working together* during visual search for a small target in naturalistic scenes (see Figure 3.8). Perhaps the most relevant

work of human-machine collaboration to ours is that of Kneusel & Mozer [111]. Such thorough study investigates the influence on human performance of the *visualization* of the intelligent system's cues used to indicate the likely target locations. Our main contribution is complementary: 1) We argue for an interaction between the humans observer performance level and that of the intelligent system in determining its influence on decisions; 2) We present eye tracking analysis to evaluate the influence of the Faster R-CNN on fixation strategies and types of errors: target not fixated (fixation errors) vs. targets fixated and missed (recognition errors).

As we will see, the general goal of this chapter is to focus on many of these questions, as there is still ongoing debate in the field of human-computer interaction regarding the use and applicability of cognitive optimizers, as well as deep learning systems supplementing human experts in the process of visual search.

3.3 Motivation for a Cognitive Optimizer

The main contribution of the first part of this chapter is the introduction and experimental verification of a real-time and feedback-enabled *attention allocation aid* (AAAD), which optimizes the operator's decision speed when they are engaging in target search, without sacrificing performance. The motivating observation is that humans have imperfect awareness of the time required to acquire all task-relevant visual information during search, and thus are generally inefficient at administering their time when scrutinizing the data sets. The proposed aid makes real-time automated search duration recommendations based on three key metrics: 1) visual search time, 2) number of eye movements executed by the observer, and 3) an estimated target detectability based on prior measurements of both the target visibility across the visual field and the observer's fixations during search. In particular, these metrics are used by the aid to estimate the time required for the operator to acquire the visual information that is necessary to support the search decisions, and subsequently indicate when this time has elapsed via a simple indicator on the user interface. We experimentally evaluate the AAAD in a simulated surveillance scenario motivated by human supervisory control, and found a factor of $\times 1.5$ increase in user efficiency in detecting and classifying targets in realistic visual imagery from a slowly moving sensor. The AAAD pipeline is generic and can readily be extended and applied to other sources of images, i.e., satellite images, astronomical images [129], x-rays for medical imaging [130, 45] or security scanning [131], and video surveillance [132] which include a human-in-the-loop [133, 134].

Our rigorous development of the AAAD also includes a number of secondary contributions, including: definition of detectability surfaces based on eye-tacking measurements, incorporation of image clutter effects, creation of a composite exploration map, and utilization of a probabilistic framework for decision making when computing overall search

satisfaction based on time, eye movements, and detectability scores.

A novel approach investigated in the current work is the design of an attention allocation aid that uses eye-tracking data to make real time inferences of the attained target detection accuracy and critically, the time to achieve asymptotic accuracy. Such estimates, which we will refer to as search satisfaction time, are utilized by the attention allocation aid to recommend that the user end the current search and move on to the next data set. In addition, if the observer completes search prior to the search satisfaction time, the eye position data can also be utilized to assess whether some area of the image remains unexplored, and suggest that the observer to further explore that area.

The success of the proposed approach requires an adequate understanding of the relation between fixational eye-movements and the accumulation of sensory evidence supporting task performance. A critical component to understanding the contribution of eye movement to task performance is the dependence of target detectability with its distance from the point of fixation, commonly referred to as *retinal eccentricity* [135, 136, 63, 46]. Indeed, this relationship can be used to build attention-based models for predicting performance [137]. Often, dynamic sensory evidence accumulation models are also dependent upon the nature of the stimuli. Our attention allocation aid relies on a set of experiments measuring how accuracy in detecting the target of interest varies with distance from fixation (retinal eccentricity) and as a function of the presentation time of the image data. These measurements are then used to implement the AAAD and validate its utility in optimizing search. To our knowledge the current approach for the AAAD and thorough experimental validation is novel to the field.

We also note that a key difference of our work in comparison to existing literature, is that our attention allocation aid is essentially a back-end search optimizer, which tells the observer *when* to stop search; rather than advising the observer *where* to look, as it does not compute fixation cue's driven by computer vision models. The effect of visual

search with such cue's driven by computer vision systems will be explored in the second part of this chapter with Faster R-CNN [9].

Humans have difficulty assessing when adequate visual information has been acquired during challenging search tasks and optimally allocating their fixations over different parts of the image [138]. The purpose of the AAAD is to utilize in real time the temporal dynamics of the eye-position data and the information acquisition process to recommend to the observers that either all information has been acquired and search can be terminated, or further exploration of the image is required. The AAAD is expected to reduce both premature image search termination and long periods of image search when no target is present without compromising the search task performance, i.e. detection and false alarm rates. Thus, the AAAD should ideally improve observer's efficiency in completing more sequential search tasks in a given allotted time period with a level of detection accuracy that is as good or better than search without the AAAD.

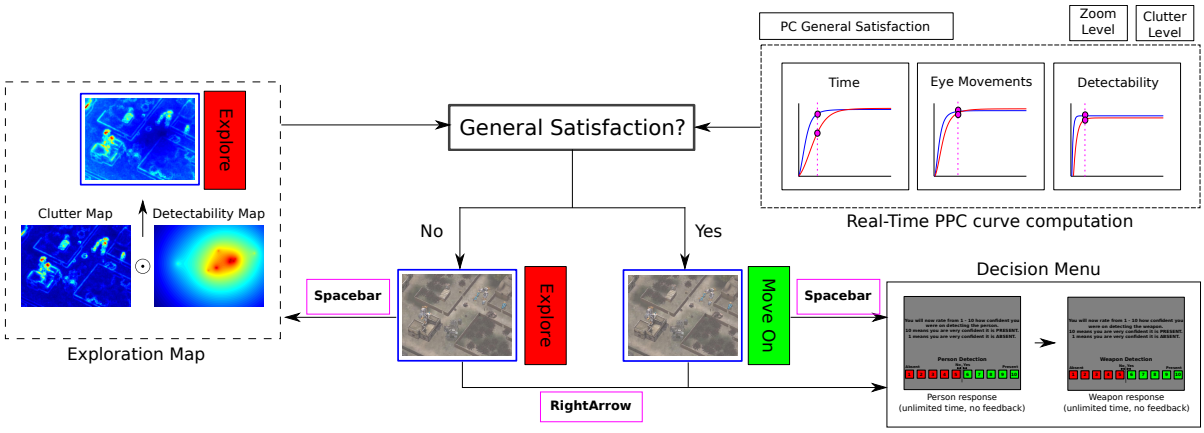


Figure 3.1: Attention Allocation Aid (AAAD) system diagram. From the start of each trial, the PPC curves are updated in real-time in the back-end, waiting for General Search Satisfaction to be achieved. Possible user inputs (highlighted in magenta) are the space bar and the right arrow. The right arrow takes the user directly to the decision menu terminating the trial, regardless of Satisfaction. The spacebar will lead the observer to the Exploration Map, and later back to the original stimuli if Satisfaction is not achieved, or to the decision menu otherwise

3.4 A Cognitive Optimizer: The Attention Allocation Aid (AAAD)

The Attention Allocation Aid system is designed to run in the background of the simulation interface (see Figure 3.1). The AAAD system starts by computing the clutter and zoom level of the input image from the slowly evolving video clip in each trial. It is assumed that the level of zoom (high, medium, or low) can be obtained from ground truth settings, given that a pilot can control a camera's zoom level, while the clutter level can be computed from the input image/video [19, 12, 10]. For our main experiments, we assumed that ground truth was provided to classify images based on clutter, since our main goal is to prove that the AAAD system works under ideal conditions[1]. A thorough investigation of the use of clutter models as ground truth predictors is left to future work.

[1]If we did *not* assume *oracle-like* inputs for zoom and clutter levels, then not finding a significant effect, could be due to poor clutter modeling, rather than poor AAAD system design.

As the trial progresses, the three PC *vs* {Time, Eye Movements, Detectability} curves (Figure 3.5) are updated in real-time using gaze location, fixation time, and saccade information obtained from the eye-tracker. The Time PPC is updated at each frame. The Eye Movements PPC is updated with an interruption-based paradigm – contingent on the eye-tracker detecting an eye movement. The Detectability PPC is updated only after every eye movement event, given that we have the fixation position and time. Each of the PC Satisfaction conditions (see Eq. (3.6, 3.7, 3.8)) is monitored independently, and once all criteria are satisfied, the AAAD systems switches from an "Explore" to a "Move On" state, where observers are encouraged to cease search, and make a decision.

Parallel to this, an *Exploration* map is computed in real-time in the back-end. The goal of the Exploration map is to inform the searcher where he/she has already searched, and to indicate the highly cluttered regions where a person is likely to be. The Exploration map has no knowledge of a target present/absent, and only uses image clutter and observer fixations.

If an observer attempts to advance to the next image while the AAAD system state is in "Explore" state, the Exploration map appears for $\sim$ 120 ms. The map is weighted by previously explored regions (computed via the detectability surface; see Supplementary Material), such that, highly cluttered and non-explored regions are highlighted. The Exploration map is computed as follows: Exploration map = FC $\odot$ (1 − Detectability Surface), where FC is the *feature congestion* [12] dense clutter map, $\odot$ is the element-wise multiplication operator, and the detectability surface is normalized to lie in the interval $[0, 1]$.

Notice that the only inputs (to the system) that the observer can produce while performing a trial (besides passively providing eye movements), are by pressing the right arrow key, which forces the trial to terminate and the subject to make a decision, irrespective of PC general satisfaction being achieved (for both AAAD, and non-AAAD

experimental sessions); or by pressing the space bar, which activates the Exploration map if PC general satisfaction is not achieved, and terminates the trial if otherwise.

3.5 Experiment 1: Psychometric Data Collection

3.5.1 Methods and Apparatus for the AAAD

Stimuli Creation: A total of 273 videos were created, each with a total duration of 120 seconds, where a 'birds eye' point-of-view camera rotated slowly around the center. While the video was in a rotating motion, there was no relative motion between any parts of the video. From a repeated subset of the original 273 videos, a total of 1440 different short clips were created, which were subsequently divided into the 4 groups (stimuli sets) that were used in subsequent experiments. Half of the clips had person present, while the other half had person absent. These short and slowly rotating clips were used instead of still images in our experiment, to simulate imagery from a moving sensor in a surveillance scenario. All clips were shown to participants in a random order. The stimuli used in all our experiments present varying levels of zoom (high, medium, low) and clutter (high, medium, low).

Apparatus: An EyeLink 1000 system (SR Research) was used to collect eye-tracking data at a frequency of 1000 Hz. Each participant sat at a distance of 76 cm from a LCD screen on gamma display, so that each pixel subtended a visual angle of $0.022\,\mathrm{deg}/\mathrm{px}$. All video clips were rendered at 1024×760 px (22.5 deg $\times$16.7 deg) and a frame rate of 24 fps. Eye movements with velocity over 22 deg/s and acceleration over 4000 deg/s^2 were qualified as saccades. Every trial began with a fixation cross, where each subject had to fixate the cross with a tolerance of 1 deg.

Experimental Setup: We performed two preliminary studies to generate the time, eye movement, and detectability PPCs: a forced fixation search (no eye movements allowed) and a free search experiment. The free search data is directly used to compute the relevant PPCs, while the forced fixation data is used to calculate detectability surfaces that allow for the computation of the detectability PPC. See Figure 3.2 for experimental

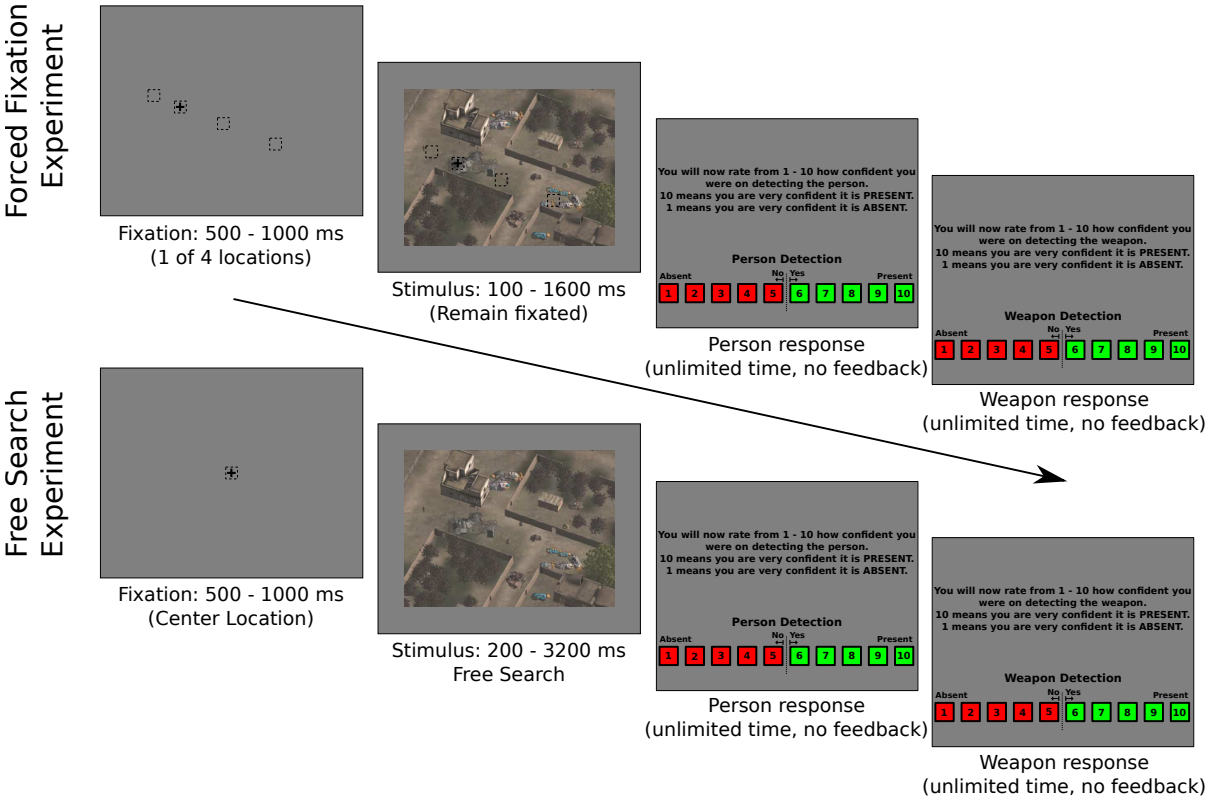


Figure 3.2: Experiment 1: The forced fixation (top) and free search (bottom) experiments to obtain the time, eye movements, and detectability PPCs.

flow.

3.5.2 Forced Fixation Search

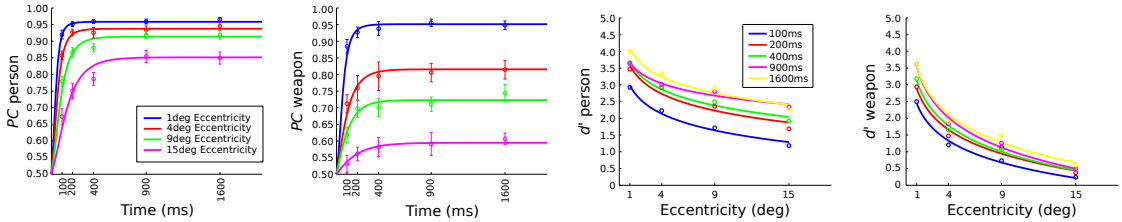
A total of 13 subjects participated in a *forced fixation search* experiment where the goal was to search within the visual periphery to identify if there was a person present or absent (yes/no task; 50% probability of person presence) and, in addition, to identify if there was a weapon present or absent (yes/no task; 50% probability of weapon presence contingent on person present). Participants had variable amounts of time (100, 200, 400, 900, 1600 ms) to view each clip. Clips were presented in a random order, with the person at a variable degree of eccentricity (1 deg, 4 deg, 9 deg, 15 deg) from point

of fixation. Subjects were not made aware of the eccentricity values used in each trial. They were then prompted with a Likert scale that required them to rate from 1-10 (by clicking on a number) their confidence of person presence. A value of 1 indicated strong confidence of person absent, and a value of 10 indicated a strong confidence of person present – intermediate values represented different levels of uncertainty. Values of 1-5 were classified as person absent, and 6-10 were classified as person present. A second rating scale (identical to the first) was then presented, requiring the subject to rate their confidence regarding weapon presence. Participants had unlimited time for making their judgments, although no subject ever took more than 10 seconds per judgment. There was no response feedback after each trial.

Each subject participated in 12 sessions that consisted of 360 clips each. There were 4 stimuli sets (each set consisted of unique images), and each participants viewed each set 3 times in random order without being aware that the images were repeated (4 sets $\times$ 3 times = 12 sessions). Every set also had the images with aerial viewpoints from different vantage points (Example: set 1 had the person at 12 o'clock – as in North, while set 2 had the person at 3 o'clock – as in East). To mitigate fixation bias, all subjects had a unique fixation point for every trial associated with each particular eccentricity value. All clips were rendered with variable levels of clutter. Each session took approximately one hour to complete. The person,i.e. search target, was of size 0.5 deg $\times$0.5 deg, 1 deg $\times$1 deg, 1.5 deg $\times$1.5 deg, depending on the zoom level. If a subject fixated outside of a 1 deg radius around the fixation cross during the trial, then the trial was aborted.

3.5.3 Free Search

A total of 11 subjects participated in a *free search* experiment where the goal was to detect and classify the person. Although eye movements were allowed, subjects were not



(a) Person and weapon detection proportion correct (PC)

(b) Person and weapon detectability curves (d')

Figure 3.3: Person and weapon detection performance in proportion correct (PC) and d' space from the forced fixation search experiment. Notice that (a) and (b) are dual representations of each other. The bottom curves in d' space will be used to generate a detectability surface.

explicitly told to foveate at the person (although they usually chose to do so). Participants had twice the amount (200 ms, 400 ms, 800 ms, 1800 ms, 3200 ms) of time than in the Forced Fixation Search. All observers began each trial with a fixation at center of the screen. They then proceeded to scan the scene to find a person and determine if the person was holding a weapon. Once the trial time was over, they were prompted with a "Person Detection" and "Weapon Detection" rating scale, and had to rate from 1-10 by clicking on a number reporting how confident they were on detecting/classifying the person. Similar to the forced fixation experiment, participants had unlimited time to make their judgments and did not receive any feedback after each trial. No trials were aborted.

Each subject in the free search experiment participated in 6 sessions that consisted of two sets of 360 unique images. In these sessions, each subject viewed one of the two sets of images, and each set was presented 3 times leading to a total of 6 sessions. Subjects were not made aware that the sessions were repeated.

3.5.4 Fitting Perceptual Performance Curves (PPCs)

Motivation of PPCs

PPCs were constructed to relate performance to each of three different metrics. The first metric is visual search time, since it is well known that time affects visual search accuracy [7] – the main intuition being that the more time a subject spends scanning an image, the higher the likelihood of detecting the target (person or weapon). The second metric is the number of eye movements a subject performs while engaging in target search. Typically, time will pass on as more eye movements are produced, but there are some cases where scrutiny in classifying or detecting a target is needed by spending long periods of fixation. As an example, one could imagine an *exploitation vs exploration* search scenario where a subject spends 1000 ms on a single fixation, given the difficulty to *classify* the target (exploitation), as opposed to a scenario where the same subject makes 3 sparse and exploratory fixations in the same 1000 ms time window to *find* the target (exploration). For this reason we chose to make time and eye movements independent metrics for our AAAD system. The third and last metric is detectability. Here, a detectability score is constructed by generating a pixel-wise map that quantifies localized information aggregation in different parts of the image (as indicated by eye movements), and subsequently combining the result to quantify the target's overall detectability. Following our previous example, one could imagine that even if an observer spends the allotted 1000 ms searching for a target and making e number of eye movements in a small spatial vicinity, it might not be a good strategy compared to spreading fixations across the image. See Figure 3.4 for an example of such fixations overlayed on different images.

We are interested in successful observer detection of the person and the weapon targets. Given that our results show that the weapon requires more time to detect

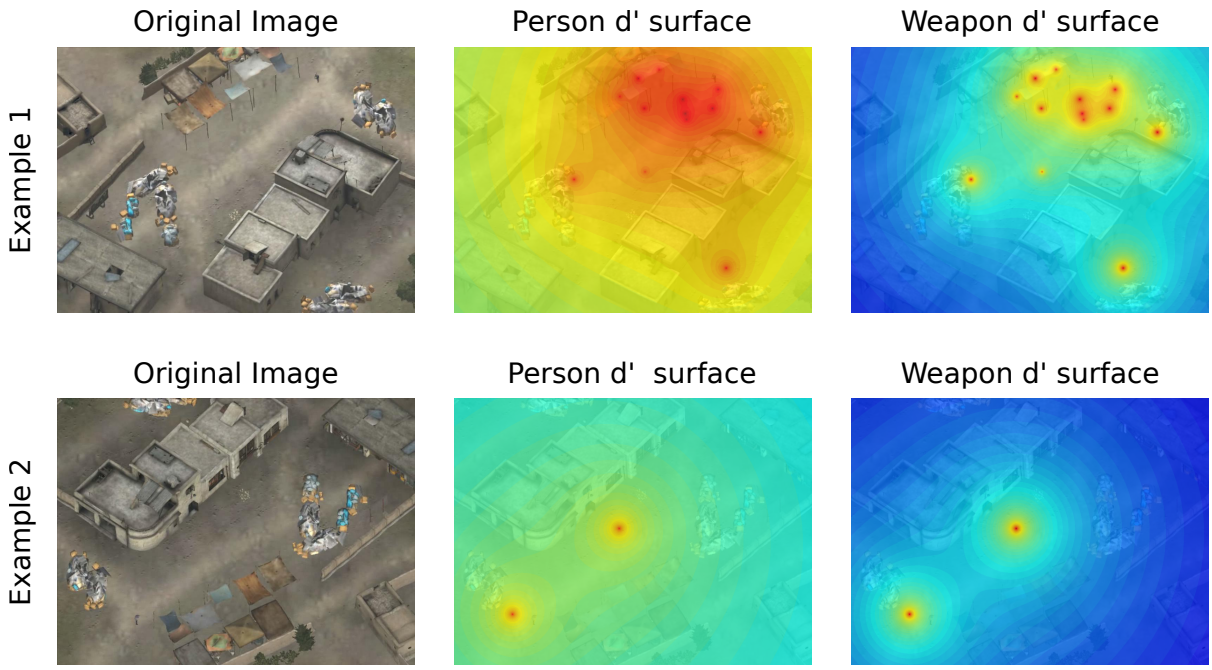


Figure 3.4: Sample person and weapon fixation maps generated from the forced fixation search experiment (Fig. 3.3(b)). These fixation maps are projections of Detectability Surfaces as described in the Supp. Mat.

than the person, the AAAD recommendation to end search was based on the PPCs corresponding to the detection of the weapon. Basing the AAAD on the PPCs for the person detection would likely compromise the detection of the weapon.

3.5.5 Computing PPCs

To model the target detection accuracy, we use the observer hit rate (the proportion of trials that the observer indicated that a target is present, given that a target is actually present in the trial stimuli) and false alarm rate (the proportion of trials that the observer indicated that the target is present, given that no target is actually pesent in the trial stimuli). Hit rates and false positive were represented as an empirical detectability index (d') and a decision criterion (λ) using an equal variance normal Signal Detection Theory (SDT) model (Green & Swets) [25]. We then fit the resulting data with curves to model

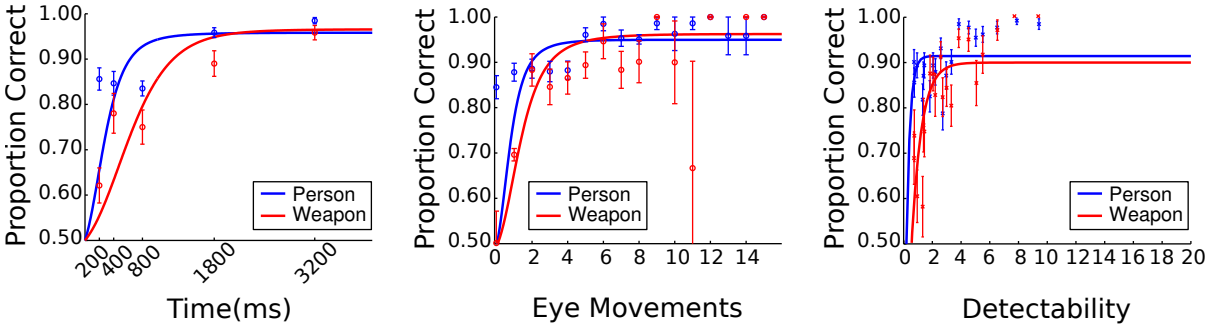


Figure 3.5: Perceptual Performance Curves (PPCs) for time (left), eye movements (center), and detectability (right) for a (low zoom, high clutter) setting. In general it takes higher time, eye movements, and detectability scores to achieve asymptotic performance for weapon detection *vs* person detection. Also shown are the error bars along each curve. Recall they have been fitted in d' space, and have been re-plotted in PC space (Eq. 3.3).

the functional relationship between the detectability indices and each of the relevant performance metrics. The best fit functions were then utilized with the equal variance SDT model to generate estimates of attained accuracy in terms of proportion correct (PC). Notice that proportion correct and hit rate are different since proportion correct takes into account both the hit rate and the correct rejection rate (proportion of trials in which the observer correctly decided that the target is absent).

For a fixed condition and setting (assuming Gaussian signal and noise distributions), the general equations to compute (d', λ) are [139]:

$$d' = Z(\text{Hit Rate}) - Z(\text{False Alarm Rate}) \tag{3.1}$$

$$\lambda = -Z(\text{False Alarm Rate}) \tag{3.2}$$

where $Z(\circ)$ is the inverse of the normal cumulative Gaussian distribution, and the hit/false alarm rates are calculated at the given experimental condition and setting. Consider as an example a condition and setting for the forced fixation search experiment:

Condition = (4 deg, 200 ms), Setting = (high zoom, low clutter). Likewise, a sample condition and setting for the free search experiment: Condition = 400 ms, Setting = (medium zoom, high clutter).

For the Time and Eye-Movements PPCs, we used a straightforward regression to find an exponential relation of the form $d'(x) = \alpha(1 - e^{-\beta x})$, where $x : \{$Time, Eye Movements$\}$ and β is constant, to obtain a continuous approximating function for the collection of points d' within a given setting.

To compute final Time and Eye Movements PPC curves, recall that there exists a function $g(\circ)$ that estimates PC, i.e., $PC(x) = g(d'(x), \lambda(x))$, where:

$$PC(x) = m(x)(\text{Hit Rate}) + n(x)(1 - (\text{False Alarm Rate})) \tag{3.3}$$

and

$$\text{False Alarm Rate}(x) = Z^{-1}(-\lambda(x)) \tag{3.4}$$

$$\text{Hit Rate}(x) = Z^{-1}(d'(x) - \lambda(x)) \tag{3.5}$$

where $m(x)$ and $n(x)$ are variables that are contingent on the number of signal (i.e. person/weapon) present and signal (i.e. person/weapon) absent trials ($m(x) + n(x) = 1, \forall x$), and the hit/false alarm rates are estimates of the true values at x. Here, curves are fitted in d' space, rather than directly from PC space, to deal with possible unbalanced datasets with signal present/absent trials (See Discussion).

In order to obtain the detectability PPC (Figure 3.5 (right)), we perform a binned regression across all trials between the respective PC performance from the free viewing task in Experiment 1 and the values of a composite detectability score D' for each image. The score D' in each trial was computed by first generating a *detectability surface* from the

user's fixations and the detectability curves (Figure. 3.3(b)), and subsequently performing a spatial average. We note that the detectability PPC is the only curve that is regressed directly to PC given that the argument of our regression function is in a d'-like space, thus $\lambda, m(x), n(x)$ need not be calculated. The training data we have for this regression is from the free search experiment.

Further details regarding the generation of the Time PPC, Eye Movements PPC, and Detectability PPC is provided in the Supplementary Material.

3.5.6 Performance Criteria and AAAD Functionality

The AAAD was designed to integrate three different inputs to compute search satisfaction. The three previous PPC inputs can be seen as individual metrics on their own, and are computed independently in the system.

Search Satisfaction Model

For computing general search satisfaction we require that all three of the following conditions are simultaneously satisfied:

$$Pr[(PC^{T}_{max} - PC(t)) < \epsilon] > \eta, \tag{3.6}$$

$$Pr[(PC^{E}_{max} - PC(e)) < \epsilon] > \eta, \tag{3.7}$$

$$Pr[(PC^{D}_{max} - PC(D')) < \epsilon] > \eta, \tag{3.8}$$

where $PC^{T}_{max}, PC^{E}_{max}, PC^{D}_{max}$ are the (fixed) asymptotic values of PC with respect to the time, number of eye-movements, and detectability PPCs given a (zoom, clutter) setting,

resp. $PC(t), PC(e), PC(D')$ are the current estimated values of PC as calculated by the time, number of eye-movements, and detectability PPCs, respectively, and ϵ, η are fixed thresholds.

An image is only said to have been adequately searched if *all* criteria equation 3.6, equation 3.7, and equation 3.8 are simultaneously satisfied. Also notice that the criteria equation 3.6, equation 3.7, and equation 3.8 are all non-decreasing in their respective arguments (which are monotonically increasing in time); thus a condition will never revert to being "unsatisfied" after being satisfied. We will refer the above search satisfaction criterion as *PC general satisfaction*.

Our motivation for using the above mentioned probabilistic framework is to take into account the error bars of each time, eye movements, and detectability curves that are zoom and clutter level dependent. We include these error bars as Gaussian standard deviations σ in our probabilistic computation:

$$Pr[(PC_{max} - PC) < \epsilon] \rightarrow 1 - Z^{-1}(\frac{PC_{max} - PC - \epsilon}{\sigma}). \tag{3.9}$$

The above strategy can be thought of centering a gaussian $(\mu_x, \sigma_x) \coloneqq (PC, \sigma)$ at every point in the PPC curves, and computing how far away the asymptotic performance PC_{max} is from every point in the curve. Thus, we will find and select the minimum point in the curve that fulfills this condition for each time, eye movements, and detectability PPC. These are the threshold PPC's that once all of them are reached in real-time on the AAAD system, the AAAD will trigger "On". A value of $\eta = 0.025, \epsilon = 0.02$ was selected for our experiments.

3.6 Experiment 2: Evaluating the Attention Allocation Aid (AAAD)

In this section we summarize a second experiment used to evaluate the effectiveness of the AAAD. Two experimental conditions were considered: A person and weapon search and classification experiment with and without AAAD. The goal of this experiment is to objectively measure any improvements in the search task performance when the subjects are assisted by the AAAD. Since it was the first time for all of our second group of subjects to participate in an eye tracking experiment, we decided to add two additional practice sessions (twenty minutes each) where we would verbally explain the non-AAAD and AAAD system.

After the practice sessions, half of the subjects were tested starting with the AAAD condition and the other half started without the AAAD condition. We counterbalanced our participants to reduce possible learning effects. The group of participants involved in Experiment 2, did not participate in and were not aware of Experiment 1. The completion of the first 2 practice sessions plus Experiment 2 with both conditions (counterbalanced) took an estimate of 2 hours for each subject. The same person and weapon present/absent statistics of Experiment 1 were used for Experiments 2 and 3. Figure 3.6 illustrates the design of the search task without (top stream) and with (bottom stream) the AAAD.

3.6.1 Condition 1: Target search without AAAD

A total of 18 subjects participated in a non-AAAD target search experiment where the goal was to complete as many trials as possible in a 20 minute interval without sacrificing task performance, where the task per trial was to correctly detect and classify the target in the minimum amount of time.

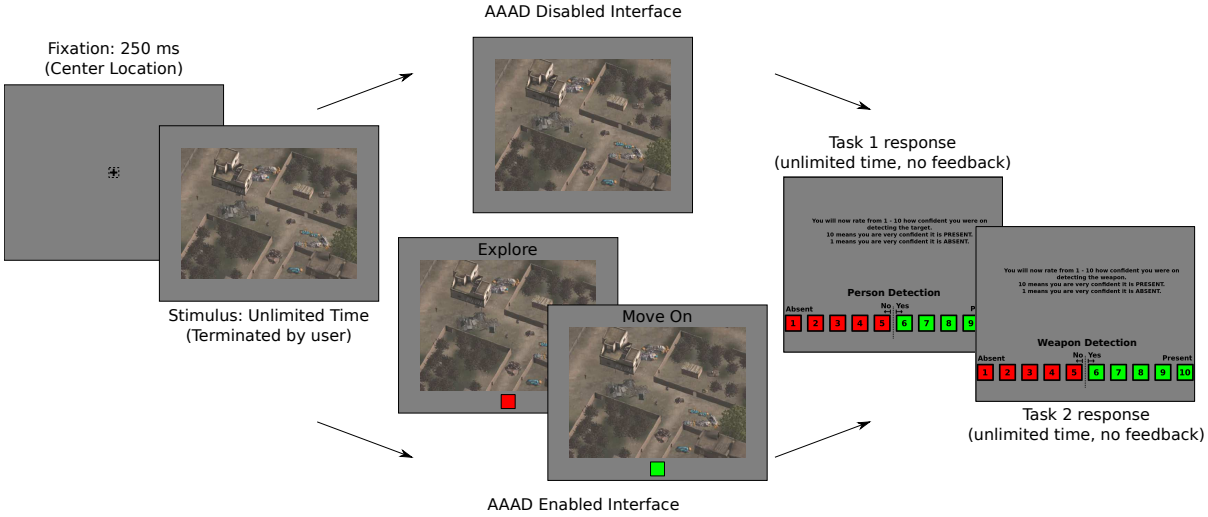


Figure 3.6: Experimental flow for Experiment 2 where we test the Attention Allocation Aid (AAAD) for condition 1 (AAAD off; top) and condition 2 (AAAD on; bottom). The top stream illustrates the non-AAAD condition, while the bottom stream illustrates the AAAD condition. Notice that the text and color in the AAAD enabled interface are coupled together: "Explore" = red, "Move On" = green. The red and green colors in the decision menus are independent of the AAAD colors.

Target detection involved reporting if the person was present or absent in the scene, and target classification involved reporting whether or not the person had a weapon. Although eye movements were allowed, subjects were not told explicitly to foveate at the target (although this was usually the case). In other words, it was possible for the subject to move on to the next trial by detecting the target in the periphery [46]. A fixation cross was placed at the center of the screen for uniform starting conditions across participants. After terminating search, subjects were prompted with a "Person Detection" rating scale where they had to rate their confidence in a person's presence on a scale from 1-10 by clicking on a number. A "Weapon Detection" rating scale then appeared where subjects also had to rate their weapon detection confidence from a scale from 1-10. Participants had unlimited time for making their judgments, though no subject took more than 10 seconds per judgment. There was no response feedback, i.e., whether their detection responses were correct after each trial.

	System Evaluation	
	Non-AAAD	**AAAD**
Average Trial Number (#)	54.33 ± 2.26	$\mathbf{57.94 \pm 1.74}$
Average Mean Time per Trial (s)	2.88 ± 0.42	$\mathbf{1.94 \pm 0.18}$
Person Hit Rate (%)	96.39 ± 1.42	$\mathbf{96.94 \pm 0.88}$
a Deep Learning Weapon Hit Rate (%)	89.37 ± 2.92	$\mathbf{92.35 \pm 2.30}$
Person False Alarm Rate (%)	2.88 ± 2.63	$\mathbf{2.42 \pm 2.04}$
Weapon False Alarm Rate (%)	8.19 ± 5.54	$\mathbf{7.15 \pm 4.70}$
Person Miss Rate (%)	3.61 ± 1.42	$\mathbf{3.06 \pm 0.88}$
Weapon Miss Rate (%)	10.63 ± 2.92	$\mathbf{7.65 \pm 2.30}$
Person Correct Rejection Rate (%)	97.12 ± 2.63	$\mathbf{97.58 \pm 2.04}$
Weapon Correct Rejection Rate (%)	91.81 ± 5.54	$\mathbf{92.85 \pm 4.70}$
Mean Trial time *vs* Time Trigger (s)	2.32 ± 0.10	$\mathbf{1.37 \pm 0.05}$
Mean Trial time *vs* EyeMvmt Trigger (s)	2.35 ± 0.10	$\mathbf{1.37 \pm 0.05}$
Mean Trial time *vs* Detect. Trigger (s)	2.70 ± 0.13	$\mathbf{1.10 \pm 0.06}$
Mean Trial time *vs* General Trigger (s)	2.70 ± 0.13	$\mathbf{1.10 \pm 0.06}$

Table 3.1: General results of the systems evaluation without and with AAAD. It should be noted that subjects were counterbalanced (half-split) to start with or without the AAAD during evaluation to compensate for learning effects. Average refers to the mean computed across observers.

3.6.2 Condition 2: Target search with AAAD

The same 18 subjects participated in a target search experiment in presence of the AAAD where the goal was same as in the previous condition. In this experiment, the AAAD was visibly turned on for the participants. They saw a text message above the center stimuli with a caption: "Explore" or "Move On", and there was a colored square below the stimuli that was colored red or green depending on the AAAD status. Participants were told to think of the AAAD as a stoplight: when it was red they should keep looking for the person/weapon, and should only move on to the next trial if the light turned green or if they were confident that they had either found the person/weapon or there was no person/weapon present.

3.6.3 Results

Table 3.1 summarizes the results of both conditions in the second experiment. There is a significant increase in number of trials per person ($M = 3.61$, $SD = 5.69$, $t(17) = 2.813$, $p = 0.015$, two-tailed), as well as a significant decrease in mean trial time ($M = -0.31$, $SD = 1.36$, $t(17) = -2.613$, $p = 0.009$,two-tailed) between AAAD conditions. Overall performance is stable, though slight improvement is seen with the AAAD. Significant differences of trial *vs* trigger times are also found. The AAAD was ran in the back-end (but not visible to the observers) in the non-AAAD condition, to compute these measures. Notice that there is a virtual speed-up factor of: $\times 1.5$, in terms of average mean time per trial across observers when using the AAAD.

In addition we performed a related samples t-test (for person t_P and weapon t_W detection) between the hit rates ($M_P = 0.54$, $SD_P = 3.65$, $t_P(17) = 0.497$, $p = 0.626$, two-tailed; $M_W = 2.97$, $SD_W = 12.42$, $t_W(17) = 1.016$, $p = 0.324$, two-tailed), false alarm rates ($M_P = -0.46$, $SD_P = 2.98$, $t_P(17) = -0.655$, $p = 0.521$, two-tailed; $M_W = -1.03$, $SD_W = 6.42$, $t_W(17) = -0.684$, $p = 0.503$, two-tailed), misses ($M_P = -0.54$, $SD_P = 4.65$, $t_P(17) = -0.497$, $p = 0.625$, two-tailed; $M_W = -2.97$, $SD_W = 12.42$, $t_W(17) = -1.016$, $p = 0.323$, two-tailed), and correct rejections ($M_P = 0.46$, $SD_P = 2.98$, $t_P(17) = 0.655$, $p = 0.521$, two-tailed; $M_W = 1.03$, $SD_W = 6.42$, $t_W(17) = 0.684$, $p = 0.503$, two-tailed), and found no significant differences between non-AAAD and AAAD conditions. This last finding is somewhat ideal as the AAAD is intended to either preserve or improve these measures.

Finally, we decided to compare the trigger times of the time, eye movements, detectability, and general satisfaction conditions of the non-AAAD sessions with the AAAD sessions. For comparison, we subtract the final trial time minus the respective trigger time. As such, these times can be thought of as offsets. Note that although the non-

AAAD condition does not show any visible assistant to the observer, the PPCs are still being computed in the back-end for the purposes of comparative data analysis. We performed four independent samples t-tests (Time: t_T, Eye Movements: t_{EM}, Detectability: t_D, General: t_G) collapsing all trials across all observers for these times and found significant differences for all trigger case scenarios, supporting the utility of the AAAD: $t_T(1920) = -8.46$, $p < 0.0001$, two-tailed; $t_{EM}(1900) = -9.03$, $p < 0.0001$, two-tailed; $t_D(1192) = -11.52$, $p < 0.0001$, two-tailed; $t_G(1190) = -11.56$, $p < 0.0001$, two-tailed.

3.7 Relevance of the Attention Allocation Aid to visual search

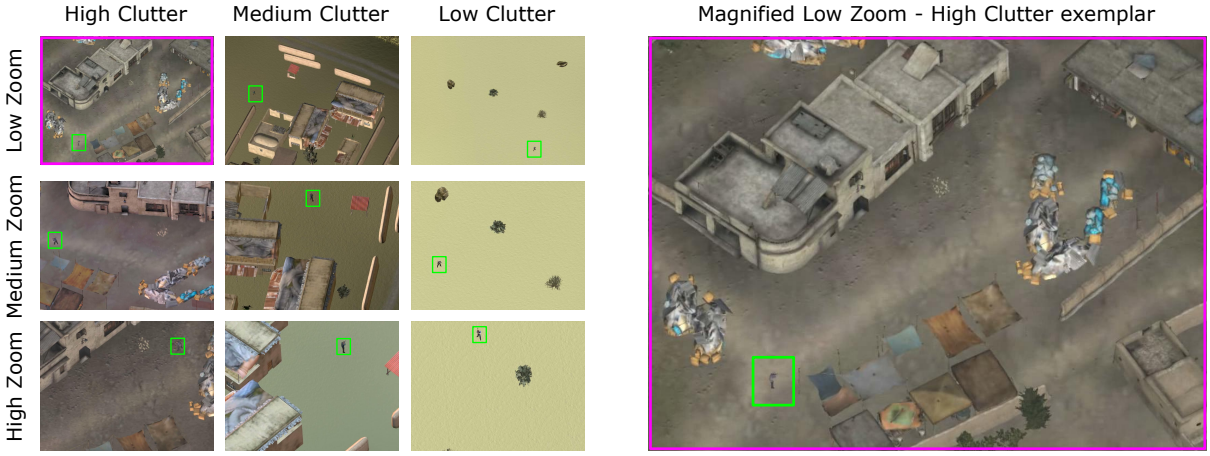
Extensions of model validity beyond current scenarios: Our results show the potential of a new approach in attention allocation aids that optimizes human search performance by utilizing real time fixational eye movements with prior measurements of target visibility across the visual field and as a function of time. However, there are various potential questions about the generalization of the model across search scenarios.

Our development of the AAAD assumed a target that is present in 50% of the images. A logical question arises as to whether the framework can generalize to real scenarios in which the target is present less frequently. Our model fits herein are performed using signal detection theory metrics (Green and Swets, 1967 [25]) that partition performance into an index of detectability, which is invariant to target prevalence, and a decision criterion, which has an optimal value (maximizing proportion correct) that varies with target prevalence. The model curves, which are utilized to make recommendations to the user, specify proportion correct as a function of time, eye movements, etc. and will vary with target prevalence. However, the model can generalize such curves to

varying estimated target prevalence assuming an optimal decision criterion for the given prevalence. Although, we have not tested the generalization experimentally, the theory accommodates such scenarios and generalizations.

For simplicity, our current work used a single target when developing the AAAD. Another natural question to ask is whether the proposed AAAD can still be used if there is the possibility of multiple targets within a given image. Indeed, it could be the case that multiple targets could change how the aid operates within a given application. In some multiple-target scenarios where the detection of *even one* target is sufficient to trigger a decision, our strategy may still apply. For example, for some medical applications, such as screening mammography, finding at least one suspicious target triggers a follow-up diagnostic mammogram. In other applications, localization of each individual targets is important and might require additional development of multiple target model curves to use in conjunction with a prior distribution of the number of targets within the images.

Impact of computer vision developments on proposed AAAD framework: The recent advanced in computer vision might seem to diminish the contributions of the proposed scheme if one assumes that all human search will eventually be replaced by machines. This is yet another reasonable question to ask, since vision has thrived in recent years, in part due to significant advances of Deep Learning [3, 140]. State of the art object recognition algorithms [75, 9, 38] have achieved high performance on certain datasets (MNIST [141], CIFAR [142], ImageNet [64]). However, the images in these datasets typically present ideal scenes with large objects at the center of an image and, currently, the ability of state of the art algorithms to find small or occluded objects in cluttered scenes (MSCOCO [143]) remains well-below that of humans. Moreover, computers often show glaring errors that humans would not make in what have been called *adversarial* examples [40] in the computer vision community (e.g. by rigging individual pixel values in an image which 'hacks' a classifier, a computer can wrongly predict that a white



(a) Sample Person Present and Weapon Present Stimuli.

Figure 3.7: Sample Stimuli of Experiments 1 and 2. Left: we show random samples of person present, and weapon present in multiple clutter and zoom conditions. Right: we show a magnified version of the (Low Zoom, High Clutter) setting. The box in green has been overlayed on each subimage to reveal the location of the person and weapon when applicable.

noise-like image is a school bus with 99% confidence [144]). Furthermore, there is still a fundamental lack of understanding with regard to the effects of computer-aided detection aids as a substitute for human observers in many application domains. For example, computer automated detection is prevalent in some countries to flag potential locations for radiologists scrutinizing x-ray mammograms. Yet there is no consensus about its contributions to improving radiologists' diagnosis accuracy (e.g. Eadie, Taylor, & Gibson, 2012 [145]). As a result of these deficiencies, human observation is still heavily relied upon in a number of applications. As a result, there are many ongoing efforts to reduce errors and optimize human visual search in life-critical tasks from military surveillance [146], to security baggage screening [147], and medical imaging [148].

What is quickly becoming prevalent across many applications is the use of a computer aid that assists humans in localizing potential targets [149]. The proposed AAAD framework does not take into account the presence of a computer aid flagging potential target locations. In some cases, the presence of a computer aid is known to guide search with the

risk of leading to over-shortened searches and missed targets that are not flagged by the computer aid [150]. The underlying model in the proposed AAAD allows calculation of an estimated observer accuracy given a pattern of fixations, time and the target visibility across the visual field. In principle, the model could be used to predict if an observer is short-cutting their search (due to the presence of the computer aid) and to alert the observer to further search the image/s. Thus, the developed AAAD framework could be potentially integrated with a computer aid, although its main contribution would likely shift from reducing search times to reducing missed targets.

Potential contribution beyond current application: Although the presented work introduces an AAAD within the context of a very specific task and images, our work serves as a proof of concept for a decision aid design approach that can potentially be applied to a variety of other applications including baggage screening and medical imaging. The model within the AAAD predicts performance on any given trial as a function of time and pattern of fixations and could be potentially used for quantifying the probability that a target was missed on a given image given the observers' search pattern. Such probabilities of error could be stored with the images and used later to identify images that require secondary inspection by an additional 2^{nd} human observer.

Arguably, the main limitation of the AAAD is that the model relies on empirically measured curves describing the detectability of the target across the visual field and as a function of time. We are currently investigating how to predict target detectability by analyzing image properties such as clutter in real time, which would greatly benefit the application of the model to broader domains.

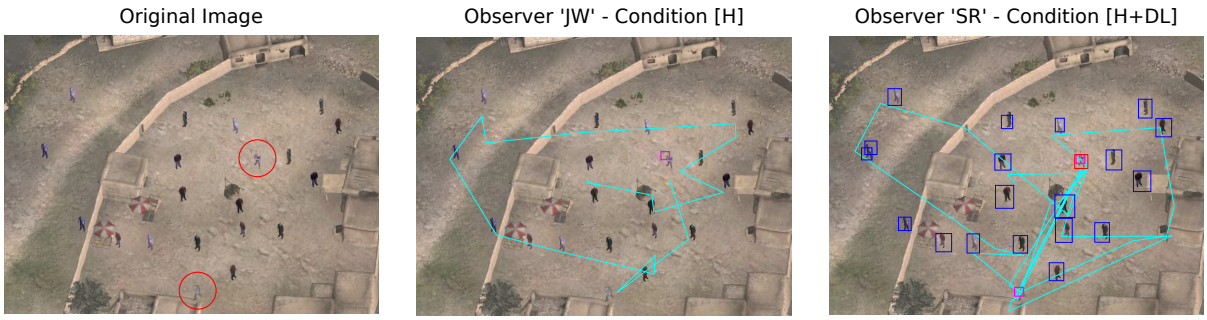


Figure 3.8: An evaluation of potential DL Benefits. Left: The original image with targets circled in red. Middle: Boxes in Magenta are clicks that observers did on target location. Right: Boxes in blue represent non-target detections and boxes in red represent target detections of the DL System. Middle and Right: Saccadic gaze pattern is plotted in cyan.

3.8 A Performance Optimizer: The Faster R-CNN object detector

We previously saw how the Attention Allocation Aid acts as a cognitive optimizer suggesting the human observer when to hault and terminate search, as no further information can be gathered. However, how would human observers perform if we constructed an aid that directly suggests where to look with an explicit cue [151]? To answer this question, we conducted a visual search experiment with a level of difficulty that is significantly greater than the previous setting, where we have a total of 20 (*vs* 1) unarmed potential targets, out of which only a sub selection of them carry weapons. Analogous to the previous experiment, the goal of the human observer is to localize the armed targets as fast as possible without sacrificing performance. Figure 3.8 show an example of such trials, where a combination of differences in eye movements, potential trial speed and detection accuracy change when multiple explicit cues of target location are computed via a deep learning expert system.

At the time of writing this chapter, one of the most popular state of the art object

detectors in computer vision which are driven by deep learning was Faster R-CNN developed by Ren et al. [9]. After the success of the AlexNet [140] Convolutional Neural Network (CNN) on the ImageNet [64] object recognition challenge, most of the field of computer vision shifted from stacking feature engineered descriptors such as SIFT [152], HOG [114] and GIST [153] with discriminative classifiers (*i.e.* Support Vector Machines, Linear Discriminative Analysis) to fully trained end-to-end models [3].

The Faster R-CNN model is thus compromised of two modules: A detection network *a.k.a.* region proposal network (RPN), and a classification network. The stacking both networks allows the system to 1) intelligently know what regions of the image to evaluate vs traditional sliding window approaches that suggest such artificial systems to look 'everywhere' in the image (even at multiple resolutions). This improvement performed by the RPN increases the speed of the network pushing the limit to near real-time localization given the $\mathcal{O}(1)$ of a single forward pass of a neural network (usually a VGG Net); 2) near optimally [2] knowing what class the bounding box or detection area belong to given the training data, and that the system has not been feature engineered (or predetermined) to know the classification boundaries for each class. Figure 3.9 shows a schematic of the Faster R-CNN object detection model.

In our experiments, we trained a Faster R-CNN object detection framework [9] which uses a VGG-Net [75] for object detection and the candidate region proposals. We picked Faster R-CNN over YOLO [39], SSD [155], R-FCN [156] given the experiments done by Huang *et al.* where they show that Faster R-CNN overperforms the other models performance-wise [157]. While running multiple object detectors in this experiment would have enriched our evaluation, we are limited by the fact that we will need multiple subjects to be ran for each DL system. One of the other reasons we did not pick YOLO over Faster

[2]Optimality here is used in the sense that the mapping from image data to class is *learned* rather than modeled, and should not be confused with optimality as in ideal observer theory [154].

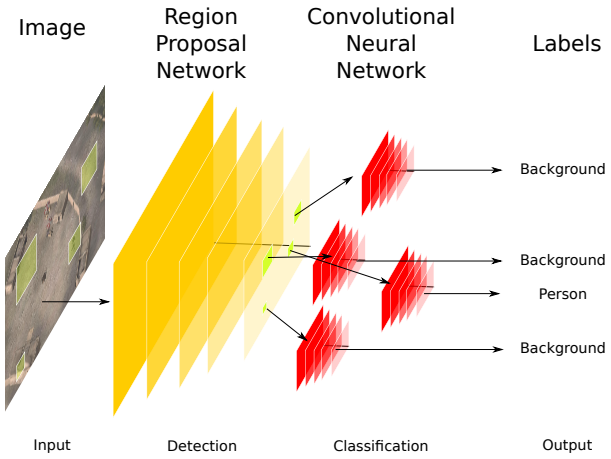


Figure 3.9: The Faster R-CNN pipeline with the Region Proposal Network and CNN detection network in the middle. the Faster R-CNN learns 'where to look' and how to evaluate candidate bounding box locations. The system is trained end-to-end.

R-CNN is that Real-Time detection in our experiments is not an issue given that we saved all the detected bounding boxes and scores in memory. In addition YOLO might not perform as well as Faster R-CNN for detecting small objects [6].

3.9 Experiment 3: Visual Search with Faster R-CNN

To analyze how man and machine work together in a visual search task, we designed an experiment with 2 main conditions: Human [H], and Human + Deep Learning [H+DL]. The search task was to find individuals holding weapons among groups of individuals without weapons. The people were embedded in complex scene. In the following subsections, we describe in detail the experiments (stimuli, experimental design & apparatus)

We evaluated the influence of the Faster R-CNN on the following human behavioral measures during visual search:

1. Target detection performance.
2. Receiver Operating Characteristic (ROC) curves.
3. Viewing time and number of trials.
4. Pattern of eye movements.

3.9.1 Creation of Stimuli

We selected 120 base images with no targets from the dataset of Deza *et al.* [106] (Experiments 1 and 2 from this chapter) that contained a variety of rendered outdoor scenes with different levels of clutter and three levels of zoom. We then randomly picked 20 locations (uniformly distributed) within each image to locate targets (individuals with weapon) and distractors (individuals without weapons). We ran a canny edge detection [158] filter to compute major edges in each images such as walls, trees and other structures. If one of the previously randomly selected locations landed on an edge, we would resample uniformly from any place in the image until a edge-less location was found. Our image generation model would also re-sample a candidate location if they

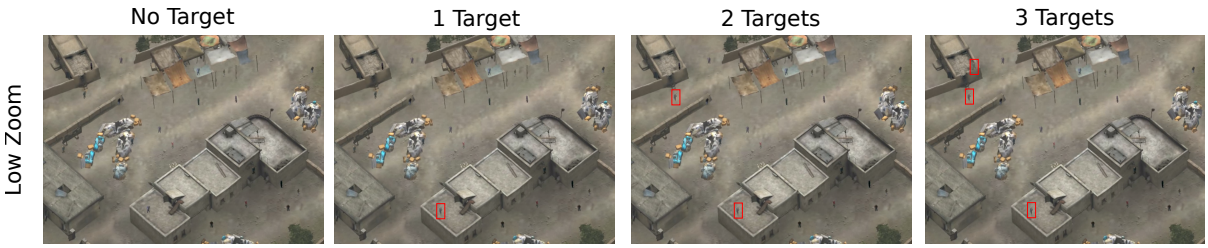


Figure 3.10: An example of a family of stimuli used in our experiment with the same image rendered with different number of targets (from left to right). The figure is better viewed when zoomed in, and illustrates the difficulty of visual search. Targets are individuals holding weapons, and they have been highlighted in red for visualization purposes.

were overlapping with a previous person location. Once the 20 locations were verified, we generated 4 different versions of the same background image such that each version had $k = \{0, 1, 2, 3\}$ targets (totalling 4×120) with the rest of candidate locations having non-targets (*a.k.a.* friends or persons without weapons). We used Poisson blending [159] on each of the locations to blend the inserted individuals into the background scene. Each image was rendered at 1024×760 px. Example scenes of the Low Zoom condition can be seen in Figure 3.10, where the difficulty of trying to find a target (a person with a weapon) is quite high.

3.9.2 Experimental Design

Our main experiment had a 2×2 factorial design to dissociate improvements caused by the DL System and those due to human learning. In the experimental design each observer participated in two consecutive sessions in one of the following orders: [H,H] (Human, Human), [H,H+DL] (Human, Human + Deep Learning), [H+DL,H] (Human + Deep Learning, Human) and [H+DL,H+DL] (Human + Deep Learning, Human + Deep Learning). Comparison of performance improvements in the Human, Human + Deep Learning vs. the Human, Human conditions allows determining whether performance

increases are due to the DL system or simply *human learning* effects. In addition, we are interested in dissecting learning and ordering effects as it could be the case that the performance differences in the second session are independent of the use of the DL system.

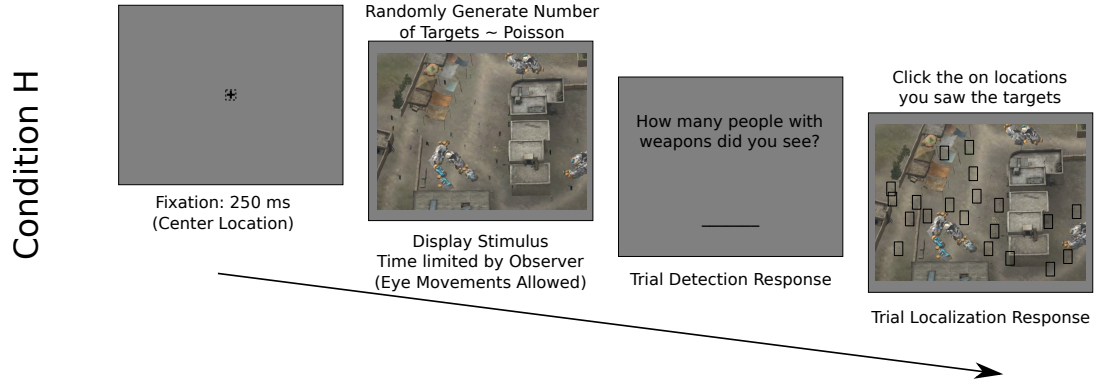
To make a direct comparison between the DL System and humans, the observers reported the number of individuals with weapons (targets). Observers also spatially localized the targets by clicking on the location of the detected target individuals on a subsequently presented image that contained the background image and bounding box locations (but no individuals) of all the potential target candidates. This evaluation paradigm is well matched to the DL system which also localizes targets with no apriori knowledge of how many targets are present in an image. The number of target per images was randomly selected with a truncated Poisson Distribution where:

$$P_k = P(X = k) = \frac{\alpha^k e^{-\alpha}}{k!} \tag{3.10}$$

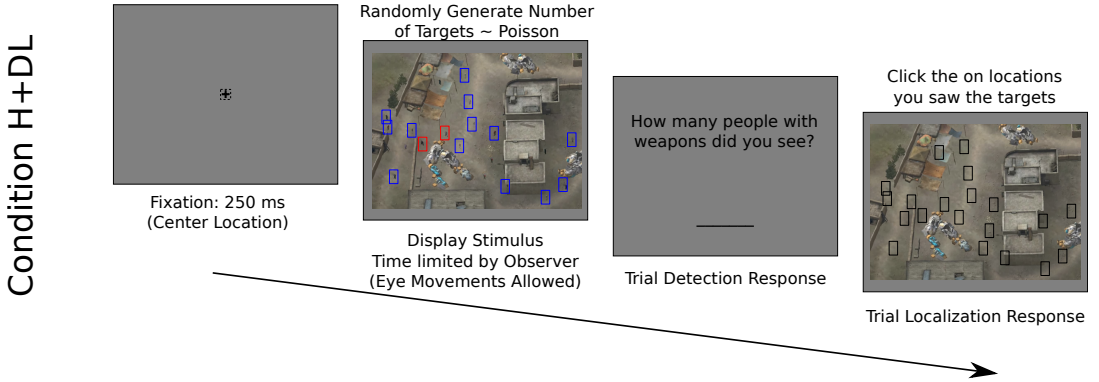
We fixed the value of $\alpha = 1$ which represents the average number of targets per trial, such that $P_0 = 0.375$; $P_1 = 0.375$; $P_2 = 0.1875$ and $P_3 = 0.0625$.

3.9.3 Apparatus

An EyeLink 1000 system (SR Research) was used to collect Eye Tracking data at a frequency of 1000Hz. Each participant was at a distance of 76 cm from a LCD screen on gamma display, so that each pixel subtended a visual angle of $0.022\,\mathrm{deg}/\mathrm{px}$. All images were rendered at 1024×760 pixels ($22.5\,\mathrm{deg} \times 16.7\,\mathrm{deg}$). Eye movements with velocity over $22\,\mathrm{deg}/s$ and acceleration over $4000\,\mathrm{deg}/s^2$ were categorized as saccades. Every trial began with a center fixation cross, where each subject had to fixate the cross with a tolerance of 1 deg.



(a) Condition [H]: Human Observer. In this condition there is no aid or cueing of targets. At the end of the trial, ground truth person locations (colored in black) are overlayed in the image to assist observers on clicking the location of potential targets.



(b) Condition [H+DL]: Human Observer + Deep Learning System. In this condition, candidate targets are cued by the DL system with color coded bounding boxes. Colors: Red is a potential foe, and Blue a potential friend.

Figure 3.11: An overview of the 2 conditions tested in the multiple target search experiment where we evaluated the benefits of a DL System in human visual search as well as the possible added benefits in terms of speed, accuracy and eye movements. Targets in these images are displayed at 0.45×0.90 d.v.a. Data was collected for conditions [H,H]; [H,H+DL]; [H+DL,H]; and [H+DL,H+DL].

3.9.4 Human: Training and Testing

A total of 120 observers divided in four groups of 30 performed the [H,H], [H,H+DL], [H+DL,H], [H+DL,H+DL] sessions respectively.

Training: Each observer engaged in 3 practice trials at the beginning of each session. Feedback was given at the end of each practice trial analogous to providing a supervised signal.

Testing: Observers were instructed to optimize two general goals: The first was to *maximize* maximize the total number of trials on each of the 20 minute sessions. The second was to *maximize* their performance when engaging in visual search. We emphasized that they had to do well maximizing both goals, such that they should not rush over the trials and do a poor job, but neither should they over dwell on search time for every image. No feedback was given at the end of each trial. See Figure 3.11 for experimental flow.

3.9.5 Deep Learning System: Training and Testing

Training: We trained the network on tensorflow [160] for over 5000 iterations as shown in Figure 3.12, after having it pre-trained with 70000 iterations on a collection of images from ImageNet achieving standard recognition performance. The images fed to the network for training were $420 = 7 \times 20 \times 3$ images, consisting of 7 rotated rotated versions and 20 person inputs (10/10 friends/foes) for each of the 3 target sizes. Small rotations, crops, mirroring and translations were used for data augmentation. The images that were rendered for testing had never been seen from the network, and were rendered with a mix of randomly sampled individuals with and without weaponse from the held out dataset.

Testing: Candidate bounding boxes developed by the system always overlayed on possible person locations irrespective of whether the individual carried a weapon. Thus the DL System never produced a Location-driven False Alarm, all mistakes delivered by the system were recognition/classification based. Bounding box candidates with a threshold lower than $\eta = 0.8$ were discarded, and overlaying bounding boxes (doubles) were removed with non-maximal suppression (NMS).

With these configurations both the DL System and the Human are prone to make

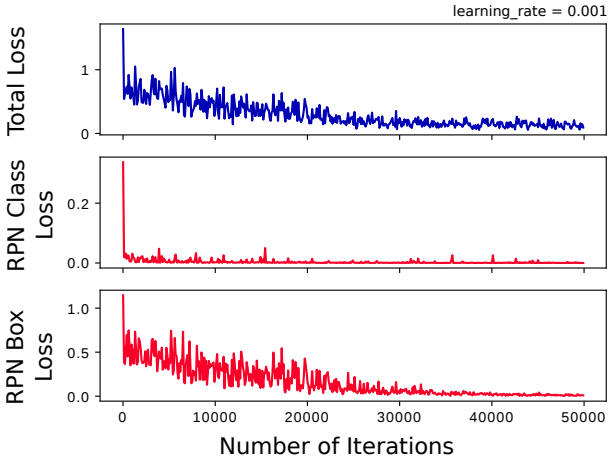


Figure 3.12: Training loss for the Faster R-CNN trained after 50k iterations. We used the model trained after 5000 iterations to avoid over-fitting. Having a relatively high performing (but not perfect) system is ideal to split observers into high and low sensitivity groups for post-hoc analysis.

the same type of judgments and mistakes. For example: 1) Humans are not allowed to click on the same locations more than twice (computer as well given NMS); 2) The Human and DL system both have a finite collection of possible locations from where to select the target locations. In addition, the experiment is free-recall for humans as they are allowed to report any number of targets per image without prior information. The DL system has the same criteria since the computation of target location via the Region Proposal Network (RPN) does not depend on any prior of the number of targets seen in the image.

3.10 Assessment of Faster R-CNN for collaborative man-machine search

The results focus on the subgroup of trials that showed *small targets* given the greater difficulty in detection for both man and machine.

Observer Sensitivity: We quantified the influence of the DL system across groups

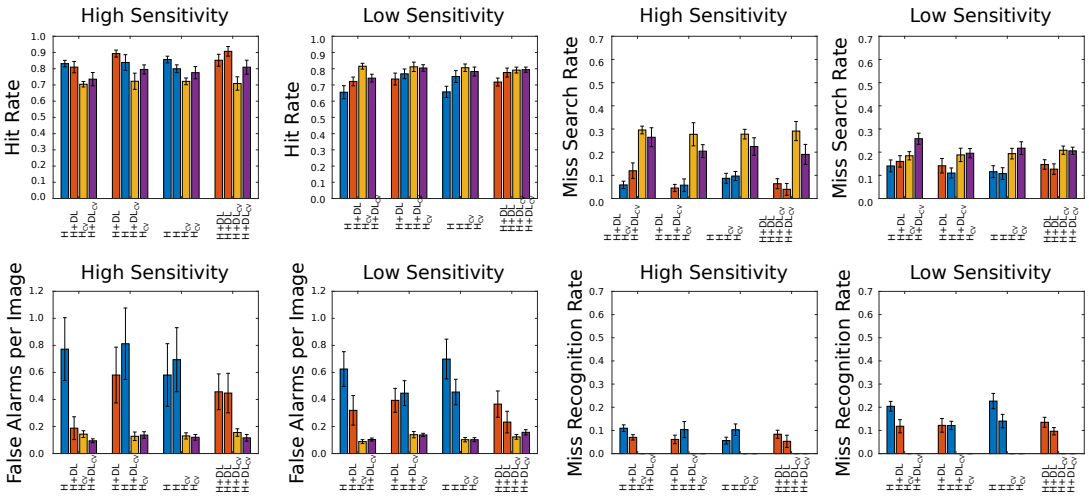


Figure 3.13: Partition of observer performance given by Sensitivity (Hit Rate) higher or lower than the machine. Hit Rate, False Alarms per Image, Miss Search Rate and Miss Recognition Rate are shown for each group. Session color code: Blue: Human without DL ; Orange: Human with DL ; Ocre: DL on 1st session; Purple: DL on 2nd session.

of observers with different abilities to find the target (hit rate). We split the participants form the [H,H+DL] condition into two groups contingent on their *sensitivity* (hit rate): the first group was the high sensitivity group who had a hit rate higher than the DL system in the first session, conversely the second group was the low sensitivity group who had a lower hit rate than the DL system. We ran an unpaired t-test to verify that there were indeed performance differences, and found a significant difference $t(27) = 3.64, p = 0.0011$ for the high sensitivity group ($M_H = 83.16 \pm 2.00\%$) and the low sensitivity group ($M_L = 65.52 \pm 4.04\%$). This effect was visible across all other conditions: [H+DL,H] with $t(28) = 3.40, p = 0.0020$, ($M_H = 89.34 \pm 2.15\%$), ($M_L = 73.66 \pm 3.67\%$); [H,H] with $t(27) = 3.96, p < 0.001$, ($M_H = 85.68 \pm 2.06\%$), ($M_L = 65.75 \pm 3.46\%$); and [H+DL,H+DL] with $t(27) = 2.21, p = 0.0351$, ($M_H = 85.24 \pm 3.68\%$), ($M_L = 71.79 \pm 2.45\%$).

3.10.1 Target Detectability

In the following subsection we describe the collection of the metrics used in our analysis that come from the signal detection theory literature [161] and medical imaging/radiology (search and recognition errors) [162]. We group such metrics contingent on the sensitivity of each observer and plot these values in Figure 3.13.

1. **Hit Rate per Image** (HR): The total number of targets correctly selected at divided by the total number of targets in the image.

2. **False Alarms per Image** (FA): The total number of false positives (disctractor individuals without weapons incorrectly labelled as targets).

3. **Miss Rate per Image** (MR): 1.0 - Hit Rate per Image. We divide the Miss Rate in two types:

    - **Search Errors Rate per Image** (SER): The total number of targets that were not foveated and missed divided by the total number of targets in the image.

    - **Recognition Errors Rate per Image** (RER): The total number of targets that were foveated, yet incorrectly perceived as friends (when they are actually foes) divided by the total number of targets in the image. It should be observed that RER and SER should add up to the Miss Rate per Image.

We performed two sets of mixed factor design ANOVA's for within conditions: [H] and [H+DL]; between conditions: order effects [H,H+DL] and [H+DL,H]; and between subjects. Each mixed ANOVA was ran separately for the high and low sensitivity groups. We found the following results:

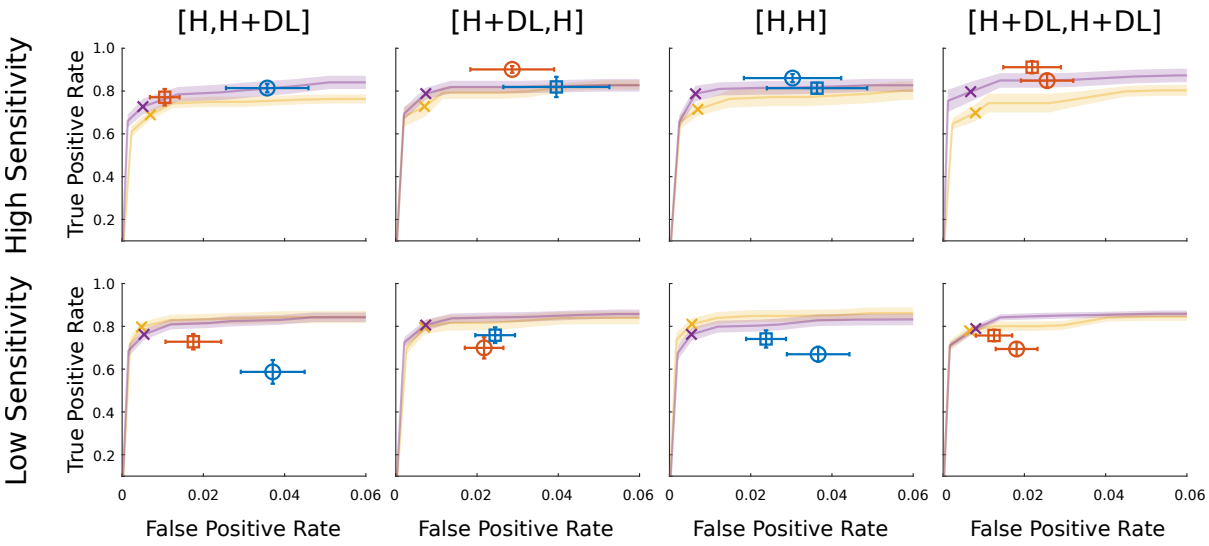


Figure 3.14: ROC plots that compare the performance of the Human and the DL system separately and working collaboratively. The plots are split by High / Low sensitivity, and Experimental Condition: [H,H+DL], [H+DL,H], [H,H] and [H+DL,H+DL]. ROC's in ocre and purple show the performance of the DL System independently for the first and second session respectively. The cross indicates the operating point along the curve at $\eta = 0.8$. For the human observer a circle is the first session, and a square the second session. Blue and orange indicate presence of the DL system when engaging in visual search.

False Alarms per Image: A main effect of *reduction* of False Alarms with the presence of the DL system for both the high and low sensitivity group: $F_H(1, 24) = 7.23, p = 0.01$, and $F_L(1, 24) = 4.93, p = 0.03$.

Search Error Rate: No significant differences in terms of search error rate between conditions. Although we did find that on average the search error rate was lower for the high sensitivity group: unpaired, two-tailed, $t(116) = -3.633, p < 0.0001$.

Recognition Error Rate: No reduction in recognition error rate for the high sensitivity group, but a marginal main effect for reduction in recognition error rate for the low sensitivity group in the presence of the DL system $F_L(1, 32) = 3.85, p = 0.058$, as well as a marginal ordering effect (showing [H+DL] or [H] first) $F_L(1, 32) = 3.96, p = 0.055$.

3.10.2 Assessment of the Human and Machine Receiving Operating Characteristics

Similar to the work of Esteva *et al.* [112], we decided to investigate how do humans perform compared to the DL system when the system performs individually along its entire receiver operating characteristic (ROC) curve, including its operation point at $\eta = 0.8$. It may be possible that we find that the DL system performs much better overall than the human observers even for the high sensitivity group, as a higher sensitivity might also imply high false alarm rates and thus less discriminability. This is an effect that can usually be explained within the context of signal detection theory [161]. If the ROC point of the human observers with or without assistance is outside of the DL ROC curve (ocre and purple for the each of the 2 sessions respectively), then we can say that the humans observers collectively perform better than the machine.

To compute the ROC curve per image we require both the TPR (True Positive Rate) and FPR (False Positive Rate) per image I. Note that FPR is not be confused with False Alarms per Image as plotted in Figure 3.13. If h is the number of hits the observer performs on the image, and f the number of false alarms restricted to the clicked bounding box locations: We will compute $TPR = h/G$, and $FPR = f/(N - G)$, where $N = 20$ is the total number of possible bounding boxes that an observer has to choose from to make a selection for target present, and G is the number of true targets there are in the image $(0, 1, 2$ or $3)$. These statistics were averaged for both the machine to plot an entire ROC curve, and for the human observers plotting the ROC points as depicted in Figure 3.14.

To analyze variability in the observers behaviour as well as decision strategies we will use estimates of target detectability (d') and decision bias (λ) s.t.

$$d' = \Phi^{-1}(TPR) - \Phi^{-1}(FPR) \tag{3.11}$$

and

$$\lambda = -\Phi^{-1}(FPR) \tag{3.12}$$

where Φ^{-1} is the inverse of the cumulative normal distribution.

In what follows of the remaining subsection we focus on comparing two types of conditions across each others along previously mentioned metrics. These are mainly: [H,H+DL] *vs* [H,H], to investigate how the observer ROC changes in the second session with the presence of the DL system, and also [H+DL,H] *vs* [H+DL,H+DL] which investigates if the observer's signal detectability and criterion change as a function discarding/continuing the DL system in the second session.

Detectability (d')**:** We performed an unpaired t-test across the second sessions comparing [H,H+DL] *vs* [H,H], and [H+DL,H] *vs* [H+DL,H+DL], and did not find any statistically significant changes in d'.

Decision bias (λ)**:** Only the high sensitivity group showed differences in bias when the DL system was removed in the second session $t(24) = 2.62, p = 0.01$. $\hat{\lambda}_{H+DL} = 2.09 \pm 0.05$ *vs* $\hat{\lambda}_{H+DL} = 1.79 \pm 0.12$ in the [H,H+DL] *vs* [H,H] condition.

We finally summarized the detectability and bias scores across all observers, pooled over both sessions, and split by sensitivity and condition [H] vs [H+DL], and compared these to the machine in Table 3.2:

	detectability (d')		bias (λ)	
	[H]	[H+DL]	[H]	[H+DL]
High	2.84 ± 0.10	3.13 ± 0.09	1.82 ± 0.05	1.95 ± 0.04
Low	2.42 ± 0.10	2.62 ± 0.08	1.83 ± 0.03	2.00 ± 0.03
DL	2.78 ± 0.04		1.96 ± 0.02	

Table 3.2: Human vs DL system performance

It is clear that when removing any learning effects of session order, that *only* human observers with high sensitivity perform better than the DL system, while the low sensi-

tivity group does not surpass individual DL system performance, even when aided with the DL system itself.

3.10.3 Analysis of Viewing Time and Number of Trials

Viewing Time: We found significant ordering effects for the high sensitivity group in viewing time spent per trial $F(1,24), p = 0.05$, but did not find any effects for the presence of the DL system. However, we did find an interaction for order and presence of the DL system $F(1,24) = 24.00, p < 0.0001$. As for the low sensitivity group we did not find an ordering effect $F(1,32) = 0.74, p = 0.40$, and rather did find a main effect in the presence of the DL system $F(1,32) = 10.56, p = 0.003$. This effect is shown in Figure 3.15 as a decrease in viewing time. In addition we found an interaction of order and presence of the DL system $F(1,32) = 5.6, p = 0.02$.

Perhaps a striking and counter-intuitive difference worth emphasizing is that the low sensitivity group spends *less* time than the high sensitivity group viewing each image when the system is on independent of order. Although this is understandable as our splits are driven by the performance of the observer on their first session independent of the presence of the DL system or not. In general, bad performing observers will very likely go over the image faster than high performing observers which are more careful when examining the image. Indeed, to account for differences in the splits, we ran an unpooled t-test to compare between all the [H+DL] sessions in the high and low sensitivity groups (across all orders) and found that the average viewing time (VT) differences were $VT_H = 14.35 \pm 1.37$ seconds, and $VT_L = 9.05 \pm 0.67$ seconds, with $t(117) = 3.84, p < 0.0001$.

Number of Trials: All the results we found for Viewing Time are analogous and statistically significant when analyzing number of trials – as the total time per session

in the experiment is constricted to 20 minutes, and both these quantities are inversely proportional to each other. Figure 3.15 shows such equivalence and how a low viewing time generally translates to a high number of trials across all conditions.

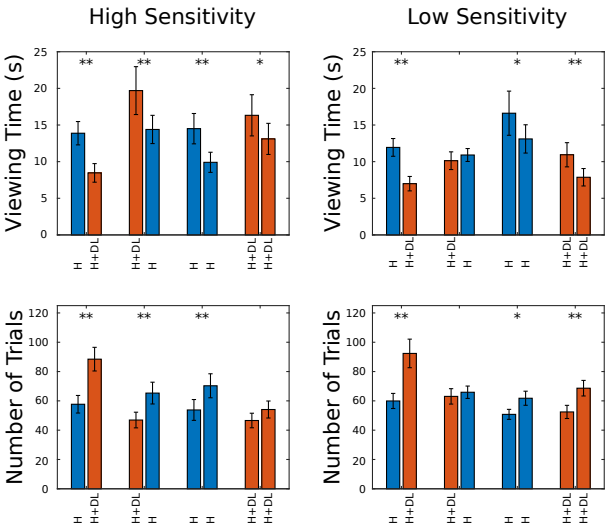


Figure 3.15: Viewing Time and Number of Trials split by high and low sensitivity observers. Blue represents the human observer [H], and orange represents the Human and Deep Learning system working together [H+DL]. 1 star represents a two-tailed independent t-test with $p < 0.05$, while 2 stars represents $p < 0.01$.

3.10.4 Analysis of Eye-Movements

Performance metrics may change as a function of the DL system as well as over each session, but how will human behaviour change as a function of such conditions? In this subsection we decided to investigate the role of eye-movements in decision making and how they may be related to performance levels. More specifically we computed the euclidean distance in degrees of visual angle (d.v.a) between the observer's fixation location f and the closest of all possible targets $\bar{t}$ as shown in Eq. 3.13:

$$D(f, \bar{t}) = \min(\bigcup_{i} ||f - t_i||) \tag{3.13}$$

To investigate such question, we decided to create boxplots of the first 5 fixations across all observers split in each one of the viewing conditions and also by sensitivity. This can be seen in Figure 3.16 which suggests that generally, observers who are enhanced when the DL system is on, fixate at a target (contingent to a target being present) by the third fixation. Thus we see how the DL system enhances fixating at the target with fewer eye movements. Qualitative and complimentary plots to this can be observed in Figure 3.17, where we show sample gaze and scan path of observers when performing search in all of these conditions.

What is most revealing about the homogeneity in fixating first at a target with the DL system on, is that this result might explain how most observers either from the high or low sensitivity group may achieve a boost in target detectability d' as shown previously in Table 3.2.

3.10.5 Main Takeaways

1. Target detection performance: The DL system reduces the False Alarm rate per Image on average across observer groups of both high/low sensitivity.

2. Receiving Operator Characteristics: We found an interaction where only the human observers with high sensitivity perform better than the DL system, while the low sensitivity group does not surpass individual DL system performance, even when aided with the DL system itself.

3. Viewing time and number of trials: The Deep Learning system only increases the number of trials for the low sensitivity group.

4. Pattern of eye movements: The DL system encourages fixating at the target by the 3rd fixation, independent of other factors.

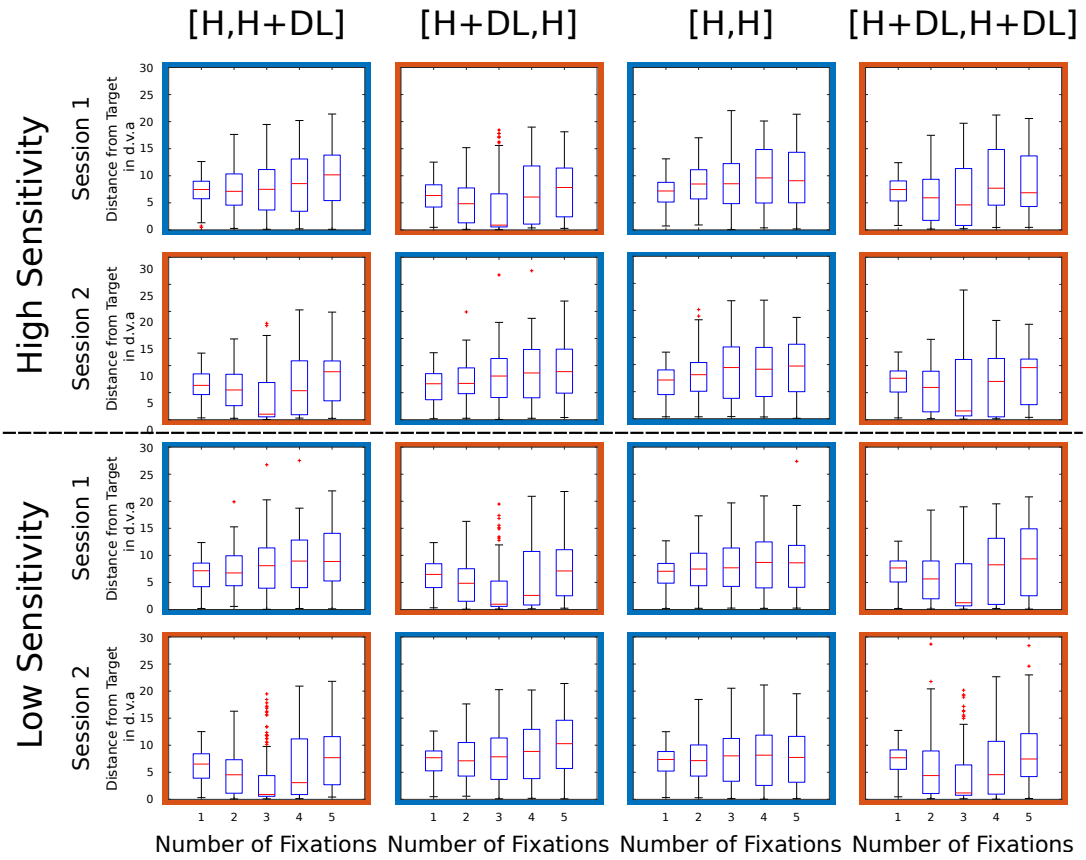


Figure 3.16: Barplots showing the 1st, 2nd and 3rd quartiles of the fixation distance to the first target foveated in degrees of visual angle (d.v.a). The Expert System aids the human by assisting him/her to fixate the target at ~ 1 deg by the 3rd fixation (orange barplots). This visual search strategy is only present when the Expert System is on – independent of the session order.

3.11 General Discussion

A central question that has remained unanswered is under what conditions do the following sets of results hold: both for the Attention Allocation Aid (AAAD), and the DL System driven by Faster R-CNN. When introducing each system, we mentioned that cognitive optimization may assist T.S.A. agents in baggage inspection, reducing long lines at the airport, but will this actually be the case? Analogously, will the Faster R-CNN be able to aid or surpass radiologists who have undergone 20+ years of training in clinical diagnosis? In essence all these tasks are different as they have different image

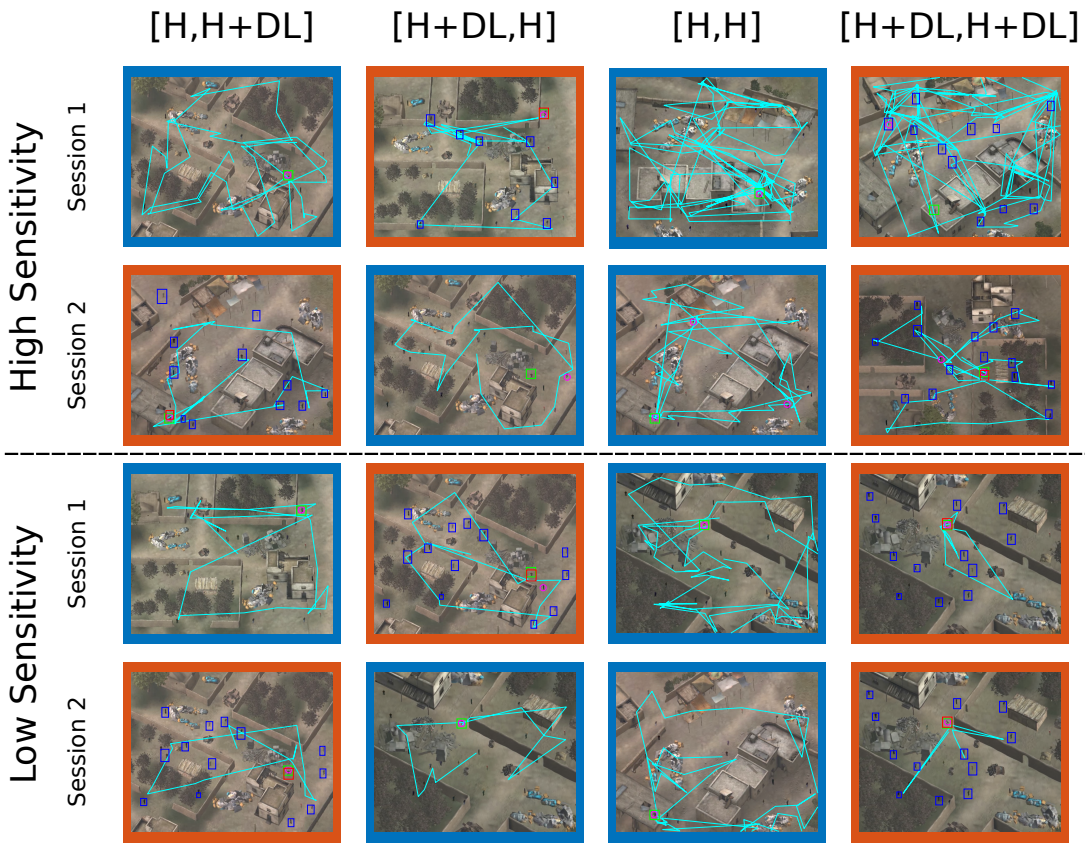


Figure 3.17: Visualization of how visual search strategies change when the DL system is on across all conditions. The lines in cyan represent the saccadic trajectories starting from the center. Boxes in blue are the DL system's detection for friend, and boxes in red are detections for targets. The box in green shows the ground truth location of the target, and circles in magenta represent the human observer's click (localization). All stimuli in this plot only have one target. Figure better viewed when zoomed in.

background and target statistics as well as performance goals; for example: it is arguably *'less worse'* to make a false alarm when diagnosing a benign tumor as a malignant one, than to mistake a water bottle for a bomb at the airport baggage clearance – though these ethical issues are open for debate. However the diversity of these problems can be reduced if we are able to directly assign computable performance values for each case scenario. Indeed, we have seen how the Faster R-CNN performs well for cases when the detectability (d') range is around 3, which is indeed quite high as the nature of this measure is non-linear. Thus, we must be formally able to compute the performance of the

human individually, as well as the machine individually to later make inferences about the potential benefits of having them work together. The analysis shown in this chapter for Experiment 3 are extendible when both human and machine perform at similar levels, and might not hold when there is stronger difference (*i.e.* $\Delta d' = 2$) between the human in the machine, where the optimal integration of the two human and machine systems under abysmal differences in performance might be excluding the low performing entity. Similarly, the locus of asymptotic thresholds developed in Experiment 2 for the AAAD is critical: if the AAAD triggers too soon, or too late, observers will ignore it.

3.11.1 The Attention Allocation Aid: A Cognitive Optimizer

Our experiments show evidence that our real-time enabled support decision system dubbed *AAAD* optimizes user efficiency in terms of an increase in the number of trials done as well as a decrease in time spent per each trial, while maintaining performance such as target hit rate and false alarm rate. Thus, the AAAD system has successfully integrated asymptotic performance of search time, eye movements and target detectability. Future analysis should of AAAD-related systems should focus on computing which of these 3 factors is the most critical when deeming asymptotic performance, *i.e.* which system triggers last. This is a critical question that we are currently exploring, and is relevant future work as it is possible that the detectability map mainly drives the system to trigger a *'Move On'* signal in contrast to the other psychometric functions that a priori we suspect are equally important, but may not be. Doing so may lead to increasing the acceleration benefits of human observers when using such cognitive optimizers.

Another factor that we did not consider in our analysis of the Attention Allocation Aid in Experiments 1 and 2, that only stemmed from Experiment 3 when evaluating the influence of Faster R-CNN on visual search, was the partitioning of human observers

into low and high sensitivity groups. Perhaps such *low*/*high* senstivity partition of observers would indeed show increases in performance beyond throughput, as the data from Table 3.1 suggests a tendency of improvement across all evaluation metrics (Hit Rate, False Alarms). It is possible that not partitioning the observers weakens the effect when statistically testing for such differences, and considering such partitions should be taken into account for future work.

Finally, in this chapter we have described how to fully implement such system through an initial set of psychometric experiments to find perceptual performance curves for target search, as well as a consequent experiment that verifies the benefits of the AAAD. Future computer-human interaction based systems could benefit from implementing AAAD-like systems where having a human-in-the-loop is critical to finding a target even beyond surveillance systems, e.g., medical imaging, astronomical data imagery and remote sensing.

3.11.2 Faster R-CNN: The Performance Optimizer

While there has been a great maturation in terms of success of deep learning systems regarding object detection, there are still many limitations in object detection, such as: adversarial examples [40], fine-grained detection [163], and small objects(targets) [164]. Adversarial examples have clearly exposed important limitations in current deep learning systems, and while having an experimental setup of visual search with and without adversarial examples would be interesting, it is not the focus of our work. The outcome is somewhat predictable and guaranteed: humans would achieve a higher recognition rate than computers – yet we do not discard the possibility that performing a study similar to ours with the presence of adversarial images is relevant and should be explored. On the other hand, future work regarding integrating human and machines in visual search

in the presence of *human-like* adversarial examples [109] might also be of great interest as explored in the recent work of Finlayson *et al.* [165] applied to medical images.

In this chapter, we thus centered our efforts in studying a more real and applicable problem which is fine-grained small object detection and classification with a limited number of training exemplars that uses a commonly deployed pre-trained VGG16 [75]. We found that, for a current DL system, its influence on human search performance interacts with the observers' sensitivity. This highlights the complexity of integration of DL systems with human experts, and it is likely that these interactions also depends on the performance level of the DL system as well as the observers' trust on the DL system. Moreover, with the recent surge of DL systems applied to Medical imaging, we believe that these experimental insights will be transferable to such and other human-machine collaborative domains.

3.11.3 Concluding Remarks for a Potential Joint-Optimization System

Some general limitations when evaluating the performance of many of these man-machine collaborative systems is the that the parameter search space to optimize the machine-based system grows, the more complex the system is. Multiple pilot studies were ran on the AAAD to determine proper asymptotic thresholds for the psychometric function, and in parallel many other studies were ran where we modified the number of training data, epochs and network architecture that the Faster R-CNN should use to perform target detection at a level that is somewhat *on par* to human performance.

However, we think the goal of designing such joint-optimization systems is still tractable. One particularity that we encountered during their development was starting from an agnostic position of now knowing if any of the single systems would improve

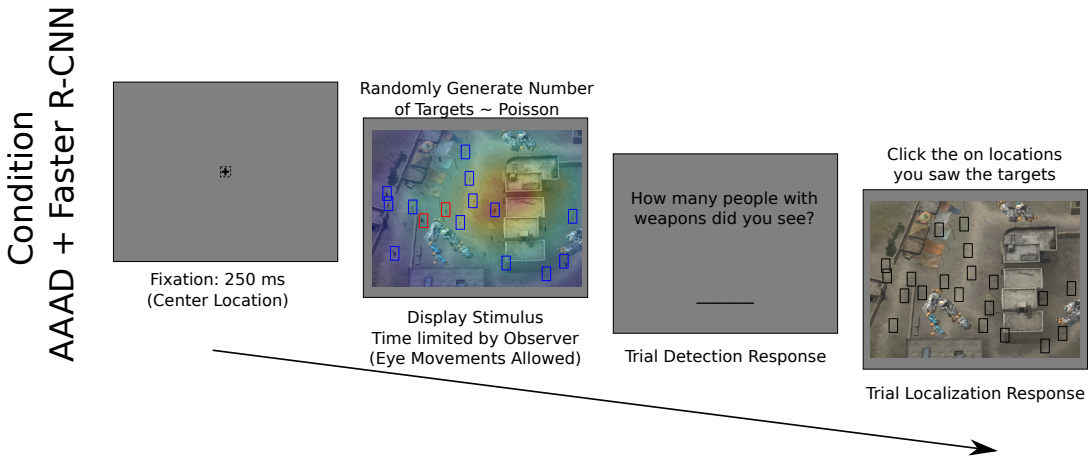


Figure 3.18: An example of a Joint-Optimization condition in an experiment where observers have both a Cognitive Optimizer such as the Attention Allocation Aid, and a Performance Optimizer such as Faster R-CNN while they engage in visual search.

both the human observer search efficiency in terms of search time, as well as their performance (computed via d' or Proportion Correct), we found that the first system dubbed the AAAD, only enhanced observer search efficiency while maintaining performance constant, and the second system – the widely known Faster R-CNN – only improved performance for the low sensitivity group while maintaining search efficiency constant.

Future work should be focused on how to integrate both systems into a single artificial agent that boosts efficiency and performance. In Figure 3.18 we show an example of such setting for our stimuli. Here an observer must perform search and use both the AAAD status and visuals, as well as the cues from the Faster R-CNN (or potentially any other object detector developed at the time). As we have shown, if the benefits from such systems are orthogonal, and potentially dependent on their group sensitivity, then constructing a single system that optimizes speed and detection may be possible.

3.12 Supplementary Material

3.12.1 Additional PPC Computation Details

Time PPC: We use the equal-variance assumption Gaussian model to retrieve λ, s.t. $\lambda(x) = -Z(\bar{f})$, where $\bar{f}$ is the average number of false alarms across all 5 time conditions (200 ms, 400 ms, 800 ms, 1800 ms, 3200 ms). We use this model because there is an equal number of person present/absent, and weapon present/absent trials (contingent on person present). This implies: $m(x) = n(x) = 0.5$.

Eye Movements PPC: Eye movements were quantized in all our experiments as the number of saccades. We estimated the observer bias λ, for every $x = 0, 1, 2, 3, ..., 15$ saccades, and later computed a weighted average (inversely proportional to the error bar size) obtaining a single estimate λ_0 to serve over all x conditions. We approximated $m(x) = m_0, n(x) = n_0$, with the constants being proportional to the average number of trials present and absent across eye movement conditions.

Detectability PPC: To create a composite detectability score (D') used as input of our Detectability PPC (Figure 3.5, right), we created a *detectability surface* as seen in Figure 3.4 based on the forced fixation detection curves. First, the forced fixation detectability curves as shown in (Fig. 3.3(b)) were obtained from the forced fixation experimental data in the dual d' space as a function of eccentricity e and parameterized by search time. A logarithmic fit of the form $d'(e) = \alpha + \beta \log(e)$, where α, β are constants, was used to produce the curve for each search time condition. Detectability was offset by 1 deg of eccentricity given the forced fixation tolerance during Experiment 1, and to avoid interpolation errors at $d'(0 \text{ deg})$.

Then, we use the curves of Fig. 3.3 to create a detectability surface as follows. We start by collecting all j fixation point locations and times: (z_j, t_j) and creating a pixel-wise mesh $\mathbf{I}$ of possible person eccentricities across the image on a per observer basis.

We compute the d' value using the curves in Fig. 3.3 for every point of the mesh $\mathbf{I}$, given the time each fixation took and the distance between the fixation location and the mesh-point location, i.e., $t = t_j$ and $z = ||z_j - z_{(p,q)}|| \forall (p,q) \in \mathbf{I}$. Note that any fixation time and eccentricity can be extrapolated from the forced fixation experiment. The surface is produced by putting a normal axis to the image plane at the location of fixation j, and performing a 3D rotation around this axis. This idea is fundamentally an adaptation of the concept of *surface of revolutions*, where the generatrix is the forced fixation function $d'(z)$ (Fig. 3.3), and the axis of rotation is perpendicular from image $\mathbf{I}$ at location z_j. We refer to this as a *single-fixation surface*, which we denote $(\text{Detectability Surface})_j$. Notice that this procedure does not require knowledge of the person location, and can be thought of a non-normalized probability map that shows likelihood of finding a person on the image given any fixation location and time. The previous computation can be easily vectorized.

Each generated surface is added linearly over each observer fixation j to compute the (multiple-fixation) detectability surface: $\text{Detectability Surface} = \sum_j (\text{Detectability Surface})_j$ over the image $\mathbf{I}$. We define the final composite detectability score D' as the spatial mean of the detectability surface over the image $\mathbf{I}$. These d' scores can be added with an $L\infty$-norm or max (single-look strategy), L1-norm (late-variability model), and L2-norm (likelihood ratio observer) [139]. We use a L1-norm since real-time computations of detectability surfaces are facilitated through vectorized addition: $O(n)$ *vs* $O(n^2)$.

3.12.2 Instruction Delivery

Delivery of instructions to participants seemed to play a role in the experiment and the use of the AAAD. In preliminary versions of the AAAD, some subjects mistakenly thought that the AAAD light (red/green) was a sign of whether the person/weapon was

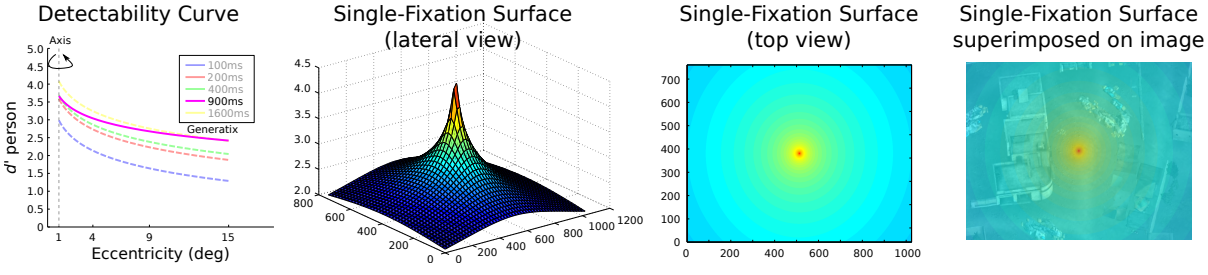


Figure 3.19: D' score generation: In this example, the observer made a single 900 ms fixation. We use the forced fixation search experiment detectability surface to get the logarithmic function that represents the decay in target detectability to compute a surface of revolutions, with the generatrix as the magenta curve, and the axis always at 1 deg on the curve, but centered at the fixation location (center of the image, in this case). The Surface is then projected in 2D on top of the image, and the final D' detectability score (used in the Detectability PPC) is the mean of this projected surface.

present or absent. In other words, they thought that the AAAD's goal was to tell them if the person/weapon was present or absent, instead of thinking of the AAAD as a search time indicator on when to terminate search. Instructions delivered in the final version (those used for the experiments herein), made this distinction clear.

The most emphasized sentence of the instructions for Experiments 2 and 3 was: *"Observers should strive to accomplish as many trials as possible without sacrificing detection performance"*. In addition, subjects were informally told: *"(...) you want to do the trials fast, but you don't want to rush and end up making careless mistakes"*.

Other details that should be taken into account for functionality of the AAAD is the possibility of subjects that were very conservative *i.e.* they only moved on when the AAAD triggered on; or, in contrast, subjects that ignored the AAAD in general *i.e.* they rarely followed the AAAD given reasons such as curiosity, possibility of deception, or general slow response times. While we cannot control for this type of behavior, this was not seen in our results subjects pool, and analyzing the data over 18 subjects is a sufficient sample size to garner trust in the overall system functionality.

3.12.3 η-Threshold Selection

High η values can lead to a very conservative thresholding for the different PPCs in the AAAD, while low η values can lead to aggressive thresholding in the multiple PPCs. Thus, finding an "optimal" value for η requires fine-tuning. To allow for this, we ran two preliminary experiments with the AAAD, (one aggressive, one conservative), to later interpolate a value that seemed reasonable, $\eta = 0.025$. Note that an aggressive η might lead to observers ignoring the AAAD, and a conservative η could be practically irrelevant to implement given its low efficiency benefits.

The optimal η value will also depend on the nature of the stimuli, the rigor of the task, and the level of expertise of the participants. For example: pilots, radiologists, security scanning personnel *vs* children, undergraduates, naive observers.

3.12.4 Exploration Map Use

The Exploration map was rarely used by the observers. Two observers did not use it at all, and one observer used the map on average at least once per trial. The mean number of times the exploration map was used per trial across all observers was 0.20 ± 0.09 requests/trial. Notice that observers observers can request the Exploration map more than once per trial.

While we were not rigorous on participant feedback, more than half the users informally reported *"I did not find the Exploration map useful, if anything I found it confusing"*, less than half of the users reported *"I used it whenever I couldn't find the target and to double-check my decisions"*. This response might be due to the short display time (120ms), which might be insufficient for visually processing the map.

3.12.5 Number of Trials Comparison

While we found that there is significance in the number of trials accomplished between the two conditions of Experiment 2, there are other factors why on average there might not be a greater difference across all participants (proportional to say the mean trial time difference). A possible reason is that some participants had smooth runs during Experiment 2 with very little or few broken fixations during the first stage of both conditions (See Figure 3.6), while other participants had more broken fixations in one Experiment or the other. Pre-trial broken fixations can be due to a subject wearing glasses, eye shape, iris color, pupil size, ethnicity, poor initial calibration, *etc.*. These are external factors that can't be controlled for, and is also why we also emphasize the significance of our results for the average trial time across subjects, which is independent of how many broken fixations they had prior to each trial, or how many trials they have accomplished.

Furthermore we performed two additional related samples t-tests to check if there were any differences in terms of response time for both tasks (target detection and classification), but did not find such differences:($M_P = 0.12, SD_P = 0.19, t_P(17) = 1.161, p = 0.262$,two-tailed;$M_W = -0.12, SD_W = 0.35, t_W(17) = -1.771, p = 0.094$,two-tailed).

3.12.6 Participant Feedback

More than one participant, informally reported *"I felt like the AAAD did not help me"*, as well as *"The AAAD helped me confirm my decisions"* and both of these opinions seemed to be spread out across the pool of participants, and did not seem to hold any relationship with their actual performance. Our most interesting feedback was given by two or three participants who explicitly mentioned that they felt like the AAAD was indirectly *pressuring* them to complete each trial before it fired on. This last feedback

is quite interesting, since it implies that behavior for certain individuals was motivated by trying to *beat* the AAAD, rather than seeing it as a complimentary aid to for search. This should be explored in future work.

Chapter 4

Conclusion

Given that human observers have a foveated visual system, we modeled the losses in the visual field in Chapter 1 by stacking a peripheral architecture on top of a clutter map, and used foveated pooling to simulate the effects of crowding done in the visual field. We found that the foveated clutter model correlated stronger with target detectability and human ratings than non-foveated clutter models when human observers engage in forced fixation search, visual search and explicit judgments respectively. This in turn lead us to ask ourselves if there is a way to visualize such perceptual losses in the visual field? In Chapter 2, we were inspired by previous work on visual metamerism and losses of information in the visual periphery as a consequence of crowding (simulated via texture-like distortions), and developed a new near real-time generative metamer model that could help us study how humans perceive the environment contingent on a point of fixation. Indeed, as humans are limited in spatial resolution when engaging in visual search, in Chapter 3 we sought out to design two aids for the human observer which provided orthogonal benefits. The first was a cognitive optimizer that is aware of the foveated nature of the human visual system and suggests where the human observer should look given global clutter maps, as well as when the observer should stop looking

as they are usually unaware of reaching satisfaction of search – thus the system increased the total image throughput. The second, was the testing of a well established deep learning expert system which specifically takes into account the information of the image and *'foveates everywhere'* unlike the human, who must be selective and plan where to make multiple eye-movements to maximize his/her efficiency. We found that this type of aid manages to increase the human observers performance mainly by decreasing false alarm rates.

Finally, there are some cases where it is possible to integrate the work in these chapters for future understanding of the visual periphery via the peripheral representations discussed in Chapters 1 and 2, along with the engineering systems developed in Chapter 3.

Integration of Foveated Clutter Models and Metamerism: The new metamer model we propose relies on a perceptual optimization process that is performed with reference to the previously rendered metamers of Freeman & Simoncelli [1]. This perceptual optimization essentially computes the maximal amount of distortions permissible for each receptive field as a function of retinal eccentricity. Future work should leverage the Peripheral Integration (PI) coefficient as a standalone reference-free metric to perform the perceptual optimization *per* receptive field. Thus future work should modify the PI such that it is bounded and between $[0, 1]$.

Integration of Foveated Clutter Models and Search Optimizers: In Chapter 3, the AAAD used a detectability (visibility) map, that was based on a system that had ideal knowledge of the global levels of clutter in the scene. A potential next step is to create an AAAD that intelligently computes the image complexity and can dynamically re-weight the target and weapon detectability curves contingent on local levels of clutter in the image, *vs* having curves that are tangentially isotropic [63]. These models that dynamically adjust target detectability likelihoods may also be extendible given the multi-

fixation integration rules that we described in Chapter 1 which extend the PI coefficient's nature from single fixation to multi-fixation search scenarios.

Integration of Metamerism and Search Optimizers: Deep Learning expert systems such as the Faster-RCNN [9] used in Chapter 3 may run on images that are dynamically re-rendered simulating the field of view of human observers contingent on observers' current fixations, giving a visual explanation of why human observers might have raised false positives or missed targets due to visual crowding around potential targets. This idea is similar to the use of Ideal Observers [154] that generally compute maximum detection performance for human observers in visual search tasks where the image and stimuli statistics are perfectly known such as Gabors embedded in spectral $1/f$ noise.

The work presented in this thesis finally suggests the design of a hybrid perceptual system – a system that integrates elements of biological and artificial visual perception – which in parallel fashion can compute and render the distortions as perceptual losses in the foveated visual field. The benefits of such hybrid perceptual systems are two-fold: they could dissect the nature of human visual search through a foveated generated simulation, and they may open the door to dynamic perception models that extend current deep learning systems that process images with a uniform resolution.

Bibliography

[1] J. Freeman and E. P. Simoncelli, *Metamers of the ventral stream*, *Nature neuroscience* **14** (2011), no. 9 1195–1201.

[2] B. Cheung, E. Weiss, and B. Olshausen, *Emergence of foveal image sampling from learning to attend in visual scenes*, *International Conference of Learning Representations (ICLR)* (2017).

[3] Y. LeCun, Y. Bengio, and G. Hinton, *Deep learning*, *Nature* **521** (2015), no. 7553 436–444.

[4] L. A. Gatys, A. S. Ecker, and M. Bethge, *Image style transfer using convolutional neural networks*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2414–2423, 2016.

[5] M. P. Eckstein, K. Koehler, L. E. Welbourne, and E. Akbas, *Humans, but not deep neural networks, often miss giant targets in scenes*, *Current Biology* **27** (2017), no. 18 2827–2832.

[6] J. Redmon and A. Farhadi, *Yolo9000: Better, faster, stronger*, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6517–6525, IEEE, 2017.

[7] M. P. Eckstein, *Visual search: A retrospective*, *Journal of Vision* **11** (2011), no. 5 14–14.

[8] R. Rosenholtz, Y. Li, J. Mansfield, and Z. Jin, *Feature congestion: a measure of display clutter*, in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 761–770, ACM, 2005.

[9] S. Ren, K. He, R. Girshick, and J. Sun, *Faster r-cnn: Towards real-time object detection with region proposal networks*, in *Advances in neural information processing systems*, pp. 91–99, 2015.

[10] A. Oliva, M. L. Mack, M. Shrestha, and A. Peeper, *Identifying the perceptual dimensions of visual complexity of scenes*, Cognitive Science Society, 2004.

[11] J. M. Henderson, M. Chanceaux, and T. J. Smith, *The influence of clutter on real-world scene search: Evidence from search efficiency and eye movements*, *Journal of Vision* **9** (2009), no. 1 32–32.

[12] R. Rosenholtz, Y. Li, and L. Nakano, *Measuring visual clutter*, *Journal of vision* **7** (2007), no. 2 17–17.

[13] R. Pramod and S. Arun, *Do computational models differ systematically from human object perception?*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1601–1609, 2016.

[14] D. M. Levi, *Crowdingan essential bottleneck for object recognition: A mini-review*, *Vision research* **48** (2008), no. 5 635–654.

[15] D. Whitney and D. M. Levi, *Visual crowding: a fundamental limit on conscious perception and object recognition*, *Trends in cognitive sciences* **15** (2011), no. 4 160.

[16] D. G. Pelli, *Crowding: A cortical constraint on object recognition*, *Current opinion in neurobiology* **18** (2008), no. 4 445–451.

[17] M. Bolduc and M. D. Levine, *A review of biologically motivated space-variant data reduction models for robotic vision*, .

[18] E. P. Simoncelli and W. T. Freeman, *The steerable pyramid: A flexible architecture for multi-scale derivative computation*, in *icip*, p. 3444, IEEE, 1995.

[19] C.-P. Yu, W.-Y. Hua, D. Samaras, and G. Zelinsky, *Modeling clutter perception using parametric proto-object partitioning*, in *Advances in Neural Information Processing Systems*, pp. 118–126, 2013.

[20] C.-P. Yu, D. Samaras, and G. J. Zelinsky, *Modeling visual clutter perception using proto-object segmentation*, *Journal of vision* **14** (2014), no. 7 4–4.

[21] R. Tudor Ionescu, B. Alexe, M. Leordeanu, M. Popescu, D. P. Papadopoulos, and V. Ferrari, *How hard can it be? estimating the difficulty of visual search in an image*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2157–2166, 2016.

[22] R. van den Berg, F. W. Cornelissen, and J. B. Roerdink, *A crowding model of visual clutter*, *Journal of Vision* **9** (2009), no. 4 24–24.

[23] M. J. Bravo and H. Farid, *A scale invariant measure of clutter*, *Journal of Vision* **8** (2008), no. 1 23–23.

[24] M. F. Asher, D. J. Tolhurst, T. Troscianko, and I. D. Gilchrist, *Regional effects of clutter on human target detection performance*, *Journal of vision* **13** (2013), no. 5 25–25.

[25] D. Green and J. Swets, *Signal detection theory and psychophysics. 1966*, *New York* **888** (1966) 889.

[26] A. Deza, G. Taylor, and M. Eckstein, *The influence of visual clutter on search guidance with complex scenes*, *Journal of Vision* **16** (2016), no. 12 1320–1320.

[27] M. S. Landy and J. R. Bergen, *Texture segregation and orientation gradient*, *Vision research* **31** (1991), no. 4 679–691.

[28] P. J. Burt and E. H. Adelson, *The laplacian pyramid as a compact image code*, *Communications, IEEE Transactions on* **31** (1983), no. 4 532–540.

[29] M.-Y. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa, *Entropy rate superpixel segmentation*, in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 2097–2104, IEEE, 2011.

[30] A. Levinshtein, A. Stere, K. N. Kutulakos, D. J. Fleet, S. J. Dickinson, and K. Siddiqi, *Turbopixels: Fast superpixels using geometric flows*, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **31** (2009), no. 12 2290–2297.

[31] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, *Slic superpixels*, tech. rep., 2010.

[32] K. Fukunaga and L. D. Hostetler, *The estimation of the gradient of a density function, with applications in pattern recognition*, *Information Theory, IEEE Transactions on* **21** (1975), no. 1 32–40.

[33] D. Comaniciu and P. Meer, *Mean shift: A robust approach toward feature space analysis*, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **24** (2002), no. 5 603–619.

[34] A. Vedaldi and B. Fulkerson, *Vlfeat: An open and portable library of computer vision algorithms*, in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1469–1472, ACM, 2010.

[35] R. Rosenholtz, *Capabilities and limitations of peripheral vision*, *Annual Review of Vision Science* **2** (2016) 437–457.

[36] R. Snowden, R. J. Snowden, P. Thompson, and T. Troscianko, *Basic vision: an introduction to visual perception*. Oxford University Press, 2012.

[37] B. Wolfe, J. Dobres, R. Rosenholtz, and B. Reimer, *More than the useful field: Considering peripheral vision in driving*, *Applied ergonomics* **65** (2017) 316–325.

[38] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, *arXiv preprint arXiv:1512.03385* (2015).

[39] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, *You only look once: Unified, real-time object detection*, *arXiv preprint arXiv:1506.02640* (2015).

[40] I. J. Goodfellow, J. Shlens, and C. Szegedy, *Explaining and harnessing adversarial examples*, *arXiv preprint arXiv:1412.6572* (2014).

[41] I. Evtimov, K. Eykholt, E. Fernandes, T. Kohno, B. Li, A. Prakash, A. Rahmati, and D. Song, *Robust physical-world attacks on machine learning models*, *arXiv preprint arXiv:1707.08945* (2017).

[42] G. Touya, C. Hoarau, and S. Christophe, *Clutter and map legibility in automated cartography: a research agenda*, *Cartographica: The International Journal for Geographic Information and Geovisualization* **51** (2016), no. 4 198–207.

[43] E. P. Simoncelli and B. A. Olshausen, *Natural image statistics and neural representation*, *Annual review of neuroscience* **24** (2001), no. 1 1193–1216.

[44] A. Oliva, *Gist of the scene*, in *Neurobiology of attention*, pp. 251–256. Elsevier, 2005.

[45] T. Drew, M. L.-H. Vo, A. Olwal, F. Jacobson, S. E. Seltzer, and J. M. Wolfe, *Scanners and drillers: Characterizing expert visual search through volumetric images*, *Journal of vision* **13** (2013), no. 10 3–3.

[46] A. Deza and M. P. Eckstein, *Can peripheral representations improve clutter metrics on complex scenes?*, in *Neural Information Processing Systems*, 2016.

[47] J. Portilla and E. P. Simoncelli, *A parametric texture model based on joint statistics of complex wavelet coefficients*, *International Journal of Computer Vision* **40** (2000), no. 1 49–70.

[48] J. Freeman, C. M. Ziemba, D. J. Heeger, E. P. Simoncelli, and J. A. Movshon, *A functional and perceptual signature of the second visual area in primates*, *Nature neuroscience* **16** (2013), no. 7 974–981.

[49] J. A. Movshon and E. P. Simoncelli, *Representation of naturalistic image structure in the primate visual cortex*, in *Cold Spring Harbor symposia on quantitative biology*, vol. 79, pp. 115–122, Cold Spring Harbor Laboratory Press, 2014.

[50] E. Akbas and M. P. Eckstein, *Object detection through exploration with a foveated visual field*, *arXiv preprint arXiv:1408.0814* (2014).

[51] L. Parkes, J. Lund, A. Angelucci, J. A. Solomon, and M. Morgan, *Compulsory averaging of crowded orientation signals in human vision*, *Nature neuroscience* **4** (2001), no. 7 739.

[52] J. Palmer, *Set-size effects in visual search: The effect of attention is independent of the stimulus for simple tasks*, *Vision research* **34** (1994), no. 13 1703–1721.

[53] J. Palmer, C. T. Ames, and D. T. Lindsey, *Measuring the effect of attention on simple visual search.*, *Journal of Experimental Psychology: Human Perception and Performance* **19** (1993), no. 1 108.

[54] M. Morgan, R. Ward, and E. Castet, *Visual search for a tilted target: Tests of spatial uncertainty models*, *The Quarterly Journal of Experimental Psychology: Section A* **51** (1998), no. 2 347–370.

[55] N. V. S. Graham, *Visual pattern analyzers*. Oxford University Press, 1989.

[56] J. H. Bertera and K. Rayner, *Eye movements and the span of the effective stimulus in visual search*, *Perception & Psychophysics* **62** (2000), no. 3 576–585.

[57] D. J. Felleman and D. E. Van, *Distributed hierarchical processing in the primate cerebral cortex.*, *Cerebral cortex (New York, NY: 1991)* **1** (1991), no. 1 1–47.

[58] D. J. Kravitz, K. S. Saleem, C. I. Baker, L. G. Ungerleider, and M. Mishkin, *The ventral visual pathway: an expanded neural framework for the processing of object quality*, *Trends in cognitive sciences* **17** (2013), no. 1 26–49.

[59] C. W. Eriksen and J. D. S. James, *Visual attention within and around the field of focal attention: A zoom lens model*, *Perception & psychophysics* **40** (1986), no. 4 225–240.

[60] M. Kwon, P. Bao, R. Millin, and B. S. Tjan, *Radial-tangential anisotropy of crowding in the early visual areas*, *Journal of Neurophysiology* **112** (2014), no. 10 2413–2422.

[61] D. M. Levi, *Visual crowding*, *Current Biology* **21** (2011), no. 18 R678–R679.

[62] R. Dubey, C. S. Soon, and P.-J. B. Hsieh, *A blurring based model of peripheral vision predicts visual search performances*, *Journal of Vision* **14** (2014), no. 10 935–935.

[63] J. Najemnik and W. S. Geisler, *Optimal eye movement strategies in visual search*, *Nature* **434** (2005), no. 7031 387–391.

[64] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et. al.*, *Imagenet large scale visual recognition challenge*, *International Journal of Computer Vision* **115** (2015), no. 3 211–252.

[65] B. Balas, L. Nakano, and R. Rosenholtz, *A summary-statistic representation in peripheral vision explains visual crowding*, *Journal of vision* **9** (2009), no. 12 13–13.

[66] J. Johnson, A. Alahi, and L. Fei-Fei, *Perceptual losses for real-time style transfer and super-resolution*, in *European Conference on Computer Vision*, pp. 694–711, Springer, 2016.

[67] J. Lubin, *A human vision system model for objective picture quality measurements*, in *Broadcasting Convention, 1997. International*, pp. 498–503, IET, 1997.

[68] S. J. Daly, *Visible differences predictor: an algorithm for the assessment of image fidelity*, in *SPIE/IS&T 1992 Symposium on Electronic Imaging: Science and Technology*, pp. 2–15, International Society for Optics and Photonics, 1992.

[69] S. Keshvari and R. Rosenholtz, *Pooling of continuous features provides a unifying account of crowding*, *Journal of vision* **16** (2016), no. 3 39–39.

[70] R. Rosenholtz, J. Huang, A. Raj, B. J. Balas, and L. Ilie, *A summary statistic representation in peripheral vision explains visual search*, *Journal of vision* **12** (2012), no. 4 14–14.

[71] L. A. Gatys, A. S. Ecker, M. Bethge, A. Hertzmann, and E. Shechtman, *Controlling perceptual factors in neural style transfer*, *arXiv preprint arXiv:1611.07865* (2016).

[72] E. Akbas and M. P. Eckstein, *Object detection through search with a foveated visual system*, *PLoS computational biology* **13** (2017), no. 10 e1005743.

[73] X. Huang and S. Belongie, *Arbitrary style transfer in real-time with adaptive instance normalization*, *arXiv preprint arXiv:1703.06868* (2017).

[74] D. Ulyanov, V. Lebedev, V. Lempitsky, *et. al.*, *Texture networks: Feed-forward synthesis of textures and stylized images*, in *Proceedings of The 33rd International Conference on Machine Learning*, pp. 1349–1357, 2016.

[75] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, *arXiv preprint arXiv:1409.1556* (2014).

[76] I. Ustyuzhaninov, W. Brendel, L. A. Gatys, and M. Bethge, *What does it take to generate natural textures?*, *International Conference on Learning Representations (ICLR)* (2017).

[77] A. J. Bell and T. J. Sejnowski, *An information-maximization approach to blind separation and blind deconvolution*, *Neural computation* **7** (1995), no. 6 1129–1159.

[78] O. J. Hénaff and E. P. Simoncelli, *Geodesics of learned representations*, *arXiv preprint arXiv:1511.06394* (2015).

[79] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, *Image-to-image translation with conditional adversarial networks*, *arXiv preprint arXiv:1611.07004* (2016).

[80] J. Ballé, V. Laparra, and E. P. Simoncelli, *End-to-end optimized image compression*, *arXiv preprint arXiv:1611.01704* (2016).

[81] T. S. Wallis, M. Bethge, and F. A. Wichmann, *Testing models of peripheral encoding using metamerism in an oddity paradigm*, *Journal of vision* **16** (2016), no. 2 4–4.

[82] V. Laparra, A. Berardino, J. Ballé, and E. Simoncelli, *Perceptually optimized image rendering.*, *Journal of the Optical Society of America. A, Optics, image science, and vision* **34** (2017), no. 9 1511.

[83] T. S. A. Wallis, C. M. Funke, A. S. Ecker, L. A. Gatys, F. A. Wichmann, and M. Bethge, *Image content is more important than bouma's law for scene metamers*, *bioRxiv* (2018) [https://www.biorxiv.org/content/early/2018/07/30/378521.full.pdf].

[84] F. A. Wichmann and N. J. Hill, *The psychometric function: I. fitting, sampling, and goodness of fit*, *Perception & psychophysics* **63** (2001), no. 8 1293–1313.

[85] B. Long, C.-P. Yu, and T. Konkle, *Mid-level visual features underlie the high-level categorical organization of the ventral stream*, *Proceedings of the National Academy of Sciences* (2018) [http://www.pnas.org/content/early/2018/08/30/1719616115.full.pdf].

[86] C. M. Ziemba, J. Freeman, J. A. Movshon, and E. P. Simoncelli, *Selectivity and tolerance for visual texture in macaque v2*, *Proceedings of the National Academy of Sciences* **113** (2016), no. 22 E3140–E3149.

[87] L. Fridman, B. Jenik, S. Keshvari, B. Reimer, C. Zetzsche, and R. Rosenholtz, *Sideeye: A generative neural network based simulator of human peripheral vision*, *arXiv preprint arXiv:1706.04568* (2017).

[88] J. Long, E. Shelhamer, and T. Darrell, *Fully convolutional networks for semantic segmentation*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.

[89] S. T. Roweis and L. K. Saul, *Nonlinear dimensionality reduction by locally linear embedding*, *science* **290** (2000), no. 5500 2323–2326.

[90] P. Tabacof and E. Valle, *Exploring the space of adversarial images*, in *2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 426–433, IEEE, 2016.

[91] A. Berardino, V. Laparra, J. Ballé, and E. Simoncelli, *Eigen-distortions of hierarchical representations*, in *Advances in neural information processing systems*, pp. 3530–3539, 2017.

[92] O. J. Hénaff, *Testing a mechanism for temporal prediction in perceptual, neural, and machine representations*. PhD thesis, Center for Neural Science, New York University, New York, NY, Sept, 2018.

[93] Z. Wang, E. P. Simoncelli, and A. C. Bovik, *Multiscale structural similarity for image quality assessment*, in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2, pp. 1398–1402, Ieee, 2003.

[94] Z. Wang and Q. Li, *Information content weighting for perceptual image quality assessment.*, *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society* **20** (2011), no. 5 1185–1198.

[95] J. R. Peters, V. Srivastava, G. S. Taylor, A. Surana, M. P. Eckstein, and F. Bullo, *Human supervisory control of robotic teams: integrating cognitive modeling with engineering design*, *IEEE Control Systems* **35** (2015), no. 6 57–80.

[96] M. L. Cummings, S. Bruni, and P. J. Mitchell, *Human supervisory control challenges in network-centric operations*, *Reviews of Human Factors and Ergonomics* **6** (2010), no. 1 34–78.

[97] T. Shanker and M. Richtel, "In new military, data overload can be deadly." The New York Times, January 16,, 2011.

[98] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, *A survey on deep learning in medical image analysis*, *Medical image analysis* **42** (2017) 60–88.

[99] T. Kooi, G. Litjens, B. van Ginneken, A. Gubern-Mérida, C. I. Sánchez, R. Mann, A. den Heeten, and N. Karssemeijer, *Large scale deep learning for computer aided detection of mammographic lesions*, *Medical image analysis* **35** (2017) 303–312.

[100] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, *et. al.*, *Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning*, *arXiv preprint arXiv:1711.05225* (2017).

[101] J. Arevalo, F. A. González, R. Ramos-Pollán, J. L. Oliveira, and M. A. G. Lopez, *Representation learning for mammography mass lesion classification with convolutional neural networks*, *Computer methods and programs in biomedicine* **127** (2016) 248–257.

[102] O. Russakovsky, L.-J. Li, and L. Fei-Fei, *Best of both worlds: human-machine collaboration for object annotation*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2121–2131, 2015.

[103] P. Y. Simard, S. Amershi, D. M. Chickering, A. E. Pelton, S. Ghorashi, C. Meek, G. Ramos, J. Suh, J. Verwey, M. Wang, *et. al.*, *Machine teaching: A new paradigm for building machine learning systems*, *arXiv preprint arXiv:1707.06742* (2017).

[104] E. Johns, O. Mac Aodha, and G. J. Brostow, *Becoming the Expert - Interactive Multi-Class Machine Teaching*, in *CVPR*, 2015.

[105] P. Chattopadhyay, D. Yadav, V. Prabhu, A. Chandrasekaran, A. Das, S. Lee, D. Batra, and D. Parikh, *Evaluating visual conversational agents via cooperative human-ai games*, *arXiv preprint arXiv:1708.05122* (2017).

[106] A. Deza, J. R. Peters, G. S. Taylor, A. Surana, and M. P. Eckstein, *Attention allocation aid for visual search*, *arXiv preprint arXiv:1701.03968* (2017).

[107] S. Branson, G. Van Horn, C. Wah, P. Perona, and S. Belongie, *The ignorant led by the blind: A hybrid human–machine vision system for fine-grained categorization*, *International Journal of Computer Vision* **108** (2014), no. 1-2 3–29.

[108] R. Geirhos, D. H. Janssen, H. H. Schütt, J. Rauber, M. Bethge, and F. A. Wichmann, *Comparing deep neural networks against humans: object recognition when the signal gets weaker*, *arXiv preprint arXiv:1706.06969* (2017).

[109] G. F. Elsayed, S. Shankar, B. Cheung, N. Papernot, A. Kurakin, I. Goodfellow, and J. Sohl-Dickstein, *Adversarial examples that fool both human and computer vision*, *arXiv preprint arXiv:1802.08195* (2018).

[110] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra, *Human attention in visual question answering: Do humans and deep networks look at the same regions?*, *Computer Vision and Image Understanding* **163** (2017) 90–100.

[111] R. T. Kneusel and M. C. Mozer, *Improving human-machine cooperative visual search with soft highlighting*, *ACM Transactions on Applied Perception (TAP)* **15** (2017), no. 1 3.

[112] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, *Dermatologist-level classification of skin cancer with deep neural networks*, *Nature* **542** (2017), no. 7639 115–118.

[113] J. Malik, P. Arbeláez, J. Carreira, K. Fragkiadaki, R. Girshick, G. Gkioxari, S. Gupta, B. Hariharan, A. Kar, and S. Tulsiani, *The three rs of computer vision: Recognition, reconstruction and reorganization*, *Pattern Recognition Letters* **72** (2016) 4–14.

[114] N. Dalal and B. Triggs, *Histograms of oriented gradients for human detection*, in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 886–893, IEEE, 2005.

[115] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, *Object detection with discriminatively trained part-based models*, *IEEE transactions on pattern analysis and machine intelligence* **32** (2010), no. 9 1627–1645.

[116] B. Cheung, E. Weiss, and B. Olshausen, *Emergence of foveal image sampling from learning to attend in visual scenes*, *arXiv preprint arXiv:1611.09430* (2016).

[117] E. Akbas and M. P. Eckstein, *Object detection through search with a foveated visual system*, *PLOS Computational Biology* **13** (10, 2017) 1–28.

[118] T. B. Sheridan, *Telerobotics, Automation, and Human Supervisory Control*. MIT press, 1992.

[119] US Air Force, *Report on technology horizons, a vision for Air Force Science And Technology during 2010–2030*, tech. rep., AF/ST-TR-10-01-PR, United States Air Force. Retrieved from http://www.af.mil/shared/media/document/AFD-100727-053.pdf, 2010.

[120] M. Pavel, G. Wang, and K. Li, *Augmented cognition: Allocation of attention*, in *System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on*, pp. 6–pp, IEEE, 2003.

[121] M. E. van Bochove, L. Van der Haegen, W. Notebaert, and T. Verguts, *Blinking predicts enhanced cognitive control*, *Cognitive, Affective, & Behavioral Neuroscience* **13** (2013), no. 2 346–354.

[122] F. Vachon and S. Tremblay, *What eye tracking can reveal about dynamic decision-making*, in *Int. Conference on Applied Human Factors and Ergonomics*, (Kraków, Poland), pp. 3820–3828, July, 2014.

[123] S. P. Marshall, *The index of cognitive activity: Measuring cognitive workload*, in *Human factors and power plants, 2002. proceedings of the 2002 IEEE 7th conference on*, pp. 7–5, IEEE, 2002.

[124] U. Ahlstrom, *Subjective workload ratings and eye movement activity measures*, .

[125] B. Donmez, P. E. Pina, and M. Cummings, *Evaluation criteria for human-automation performance metrics*, in *Performance Evaluation and Benchmarking of Intelligent Systems*, pp. 21–40. Springer, 2009.

[126] J. R. Peters and L. F. Bertuccelli, *Robust task scheduling for multi-operator supervisory control missions*, *AIAA Journal on Aerospace Information Systems* (2016). To Appear.

[127] D. C. Klein, *Using Adaptive Automation to Increase Operator Performance and Decrease Stress in a Satellite Operations Environment*. PhD thesis, Colorado Technical University, 2014.

[128] M. W. Scerbo, *Adaptive automation*, in *Neuroergonomics: The Brain At Work* (R. Parasuraman and M. Rizzo, eds.), pp. 239–252. 2001.

[129] K. L. Bouman, M. D. Johnson, D. Zoran, V. L. Fish, S. S. Doeleman, and W. T. Freeman, *Computational imaging for vlbi image reconstruction*, *arXiv preprint arXiv:1512.01413* (2015).

[130] C. K. Abbey, F. W. Samuelson, A. Wunderlich, L. M. Popescu, M. P. Eckstein, and J. M. Boone, *Approximate maximum likelihood estimation of scanning observer templates*, in *SPIE Medical Imaging*, pp. 94160O–94160O, International Society for Optics and Photonics, 2015.

[131] A. T. Biggs and S. R. Mitroff, *Improving the efficacy of security screening tasks: A review of visual search challenges and ways to mitigate their adverse effects*, *Applied Cognitive Psychology* **29** (2015), no. 1 142–148.

[132] K. Shanmuga Vadivel, T. Ngo, M. Eckstein, and B. Manjunath, *Eye tracking assisted extraction of attentionally important objects from videos*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3241–3250, 2015.

[133] C. Vondrick and D. Ramanan, *Video annotation and tracking with active learning*, in *NIPS*, 2011.

[134] C. Vondrick, D. Patterson, and D. Ramanan, *Efficiently scaling up crowdsourced video annotation*, *International Journal of Computer Vision* **101** (2013), no. 1 184–204.

[135] I. Diaz, M. P. Eckstein, A. Luyet, P. Bize, and F. O. Bochud, *Measurements of the detectability of hepatic hypovascular metastases as a function of retinal eccentricity in ct images*, in *SPIE Medical Imaging*, pp. 83180J–83180J, International Society for Optics and Photonics, 2012.

[136] E. Peli, *Contrast in complex images*, *JOSA A* **7** (1990), no. 10 2032–2040.

[137] I. Krajbich and A. Rangel, *Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions*, .

[138] J. M. Wolfe, *When do i quit? the search termination problem in visual search*, in *The influence of attention, learning, and motivation on visual search*, pp. 183–208. Springer, 2012.

[139] T. D. Wickens, *Elementary signal detection theory*. Oxford university press, 2001.

[140] A. Krizhevsky, I. Sutskever, and G. E. Hinton, *Imagenet classification with deep convolutional neural networks*, in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[141] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, *Gradient-based learning applied to document recognition*, *Proceedings of the IEEE* **86** (1998), no. 11 2278–2324.

[142] A. Krizhevsky, *Learning multiple layers of features from tiny images*, .

[143] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, *Microsoft coco: Common objects in context*, in *European Conference on Computer Vision*, pp. 740–755, Springer, 2014.

[144] A. Nguyen, J. Yosinski, and J. Clune, *Deep neural networks are easily fooled: High confidence predictions for unrecognizable images*, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 427–436, IEEE, 2015.

[145] L. H. Eadie, P. Taylor, and A. P. Gibson, *A systematic review of computer-assisted diagnosis in diagnostic cancer imaging*, *European journal of radiology* **81** (2012), no. 1 e70–e76.

[146] S. A. C. Air, Land, *Fm 3-04.15, tactics, techniques, and procedures for the tactical employment of unmanned aircraft systems (uas)*, .

[147] K. S. Hale, K. Del Giudice, J. Flint, D. P. Wilson, K. Muse, and B. Kudrick, *Designing, developing, and validating an adaptive visual search training platform*, in *International Conference on Augmented Cognition*, pp. 735–744, Springer, 2015.

[148] K. Alexander, *Reducing error in radiographic interpretation*, *The Canadian Veterinary Journal* **51** (2010), no. 5 533.

[149] M. L. Giger, *Medical imaging and computers in the diagnosis of breast cancer*, in *SPIE Optical Engineering+ Applications*, pp. 918908–918908, International Society for Optics and Photonics, 2014.

[150] D. Lyell and E. Coiera, *Automation bias and verification complexity: a systematic review*, *Journal of the American Medical Informatics Association* (2016) ocw105.

[151] M. I. Posner, *Orienting of attention*, *Quarterly journal of experimental psychology* **32** (1980), no. 1 3–25.

[152] D. G. Lowe, *Distinctive image features from scale-invariant keypoints*, *International journal of computer vision* **60** (2004), no. 2 91–110.

[153] A. Oliva and A. Torralba, *Modeling the shape of the scene: A holistic representation of the spatial envelope*, *International journal of computer vision* **42** (2001), no. 3 145–175.

[154] W. S. Geisler, *Ideal observer analysis*, *The visual neurosciences* **10** (2003), no. 7 12–12.

[155] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, *Ssd: Single shot multibox detector*, in *European conference on computer vision*, pp. 21–37, Springer, 2016.

[156] J. Dai, Y. Li, K. He, and J. Sun, *R-fcn: Object detection via region-based fully convolutional networks*, in *Advances in neural information processing systems*, pp. 379–387, 2016.

[157] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, *et. al.*, *Speed/accuracy trade-offs for modern convolutional object detectors*, .

[158] J. Canny, *A computational approach to edge detection*, in *Readings in Computer Vision*, pp. 184–203. Elsevier, 1987.

[159] P. Pérez, M. Gangnet, and A. Blake, *Poisson image editing*, *ACM Transactions on graphics (TOG)* **22** (2003), no. 3 313–318.

[160] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et. al.*, *Tensorflow: A system for large-scale machine learning.*, .

[161] J. GREEN Dand SWETS, *Signal detection theory and psychophysics*, 1988.

[162] E. A. Krupinski, *Current perspectives in medical image perception*, *Attention, Perception, & Psychophysics* **72** (2010), no. 5 1205–1217.

[163] B. Hariharan, P. Arbelaez, R. Girshick, and J. Malik, *Object instance segmentation and fine-grained localization using hypercolumns*, *IEEE transactions on pattern analysis and machine intelligence* **39** (2017), no. 4 627–639.

[164] C. Eggert, D. Zecha, S. Brehm, and R. Lienhart, *Improving small object proposals for company logo detection*, in *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pp. 167–174, ACM, 2017.

[165] S. G. Finlayson, I. S. Kohane, and A. L. Beam, *Adversarial attacks against medical deep learning systems*, *arXiv preprint arXiv:1804.05296* (2018).